Summer 8-15-2018

Comparing Attributional and Relational Similarity as a Means to Identify Clinically Relevant Drug-gene Relationships

Safa Fathiamini

Comparing Attributional and Relational Similarity as a Means to Identify Clinically Relevant Drug-gene Relationships

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

By

Safa Fathiamini, M.D., M.S.

University of Texas Health Science Center at Houston

2018

Dissertation Committee:

Trevor Cohen, MBChB, PhD[1], Advisor
Elmer V Bernstam, MD, MSE[1]
Cui Tao, PhD[1]
Funda Meric-Bernstam, MD[2]

[1]The School of Biomedical Informatics
[2]MD Anderson Cancer Center

Dedication

To my mother who has always been a source of inspiration and courage for me. To the memory of my father who always believed in me. To Kian Sam and Lily Kate who are my pride and joy. And finally, to Nooria who has loved me unconditionally, and supported me throughout this, and other journeys of my life.

Acknowledgements

My deepest gratitude to my committee chair, Dr. Trevor Cohen who continuously provided assistance and guidance for my research and thesis. I would like to thank my committee member, Dr. Elmer V Bernstam who supported me with various aspects of this project. I thank Dr. Cui Tao, who always provided invaluable guidance and support. Special thanks to Dr. Funda Meric-Bernstam who supported my research, and introduced me to the team of scientists at the IPCT, MD Anderson Cancer Center, who made this work possible. Your work is amazing and has a long lasting effect on our battle with cancer.

Special thanks to Dr. Dean Sittig, who inspired and supported me on several occasions throughout my Master's and PhD program. My strong appreciation to Dr. Parsa Mirhaji, my first academic advisor, and my dear friend, who made this new chapter of my life possible by introducing me to the science and art of biomedical informatics, and supporting me with my admission to the program.

Special thanks to all the SBMI faculty and staff, Dr. Hua Xu, Dr. Todd Johnson, Alejandro Araya, Ronald Campbell, Susan Guerrero, David Ha, Dr. Jim Langabeer, and all others who have always been there for me.

Abstract

In emerging domains, such as precision oncology, knowledge extracted from explicit assertions may be insufficient to identify relationships of interest. One solution to this problem involves drawing inference on the basis of similarity. Computational methods have been developed to estimate the semantic similarity and relatedness between terms and relationships that are distributed across corpora of literature such as Medline abstracts and other forms of human readable text. Most research on distributional similarity has focused on the notion of *attributional similarity*, which estimates the similarity between entities based on the contexts in which they occur across a large corpus. A relatively under-researched area concerns *relational similarity*, in which the similarity between pairs of entities is estimated from the contexts in which these entity pairs occur together. While it seems intuitive that models capturing the structure of the relationships between entities might mediate the identification of biologically important relationships, there is to date no comparison of the relative utility of attributional and relational models for this purpose.

In this research, I compare the performance of a range of relational and attributional similarity methods, on the task of identifying drugs that may be therapeutically useful in the context of particular aberrant genes, as identified by a team of human experts. My hypothesis is that relational similarity will be of greater utility than attributional similarity

as a means to identify biological relationships that may provide answers to clinical questions, (such as “which drugs INHIBIT gene x”?) in the context of rapidly evolving domains.

My results show that models based on relational similarity outperformed models based on attributional similarity on this task. As the methods explained in this research can be applied to identify any sort of relationship for which cue pairs exist, my results suggest that relational similarity may be a suitable approach to apply to other biomedical problems. Furthermore, I found models based on neural word embeddings (NWE) to be particularly useful for this task, given their higher performance than Random Indexing-based models, and significantly less computational effort needed to create them. NWE methods (such as those produced by the popular word2vec tool) are a relatively recent development in the domain of distributional semantics, and are considered by many as the state-of-the-art when it comes to semantic language modeling. However, their application in identifying biologically important relationships from Medline in general, and specifically, in the domain of precision oncology has not been well studied.

The results of this research can guide the design and implementation of biomedical question answering and other relationship extraction applications for precision medicine, precision oncology and other similar domains, where there is rapid emergence of novel knowledge. The methods developed and evaluated in this project can help NLP applications provide more accurate results by leveraging corpus based methods that are by design scalable and robust.

Vita

1996 ..... Doctorate, Medicine, Tehran, Iran

2009 ..... Master, Information Technology, Central Queensland University, Australia

2012 ..... Master, Health Informatics, University of Texas, School of Biomedical Informatics

Publications

DelliFraine, J., Langabeer, J., Segrest, W., Fowler, R., King, R., Moyer, P., … Jollis, J. (2013). Developing an ST-elevation myocardial infarction system of care in Dallas County. American Heart Journal, 165(6), 926–931. https://doi.org/10.1016/j.ahj.2013.02.005

Fathiamini, S., Johnson, A. M., Zeng, J., Araya, A., Holla, V., Bailey, A. M., … Cohen, T. (2016). Automated identification of molecular effects of drugs (AIMED). Journal of the American Medical Informatics Association, 23(4), 758–765. https://doi.org/10.1093/jamia/ocw030

McCoy, A. B., Wright, A., Rogith, D., Fathiamini, S., Ottenbacher, A. J., & Sittig, D. F. (2014). Development of a clinician reputation metric to identify appropriate problem-medication pairs in a crowdsourced knowledge base. Journal of Biomedical Informatics, 48, 66–72. https://doi.org/10.1016/j.jbi.2013.11.010

Field of Study

Biomedical Informatics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

Biomedical literature is growing rapidly. In 2015 alone, more than 870,000 publications were added to, and indexed in Medline (*Figure 1*).("MEDLINE Citation Counts by Year of Publication," n.d.). Clinicians and other researchers that look for specific answers to

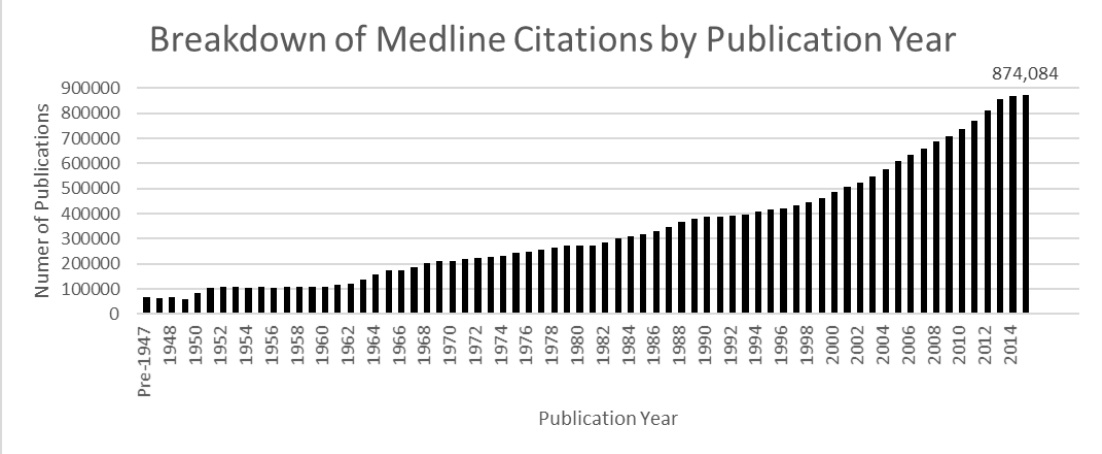

Figure 1. Medline citations by year.

their questions may be faced with overwhelmingly large sets of documents returned by information retrieval systems, such as PubMed. System that extract specific relationships

from text (such as Question Answering - QA systems) rather than documents that may contain the relationships of interest have the potential to address this problem. However, the majority of those systems rely on well-established knowledge resources (such as known relations between concepts (At, 1989)) to extract information from the biomedical literature. (Athenikos & Han, 2010) Rapidly evolving domains (such as precision oncology) pose unique challenges to QA and other relationship extraction systems. Due to the rapid emergence of new knowledge in these domains (such as discovery of new drugs or new molecular targets), the resources found in the clinical literature are scarce by definition, and systems such as SemRep, a Natural Language Processing (NLP) system for biomedical literature, which are optimized for precision, and rely solely on knowledge extracted from explicit assertions (such as "rapamycin inhibits mtor") may miss relationships of interest. (Fathiamini et al., 2016)

It has been argued that methods that infer relationships between biomedical concepts by examining the ways in which they are distributed across large text corpora, present a robust and desirable alternative (Percha & Altman, 2015). In these approaches, generally known as methods of "distributional semantics", similar representations are generated for terms that occur in similar contexts in the literature, (Trevor Cohen & Widdows, 2009) and the similarity between concepts of interest can be *measured*.

Most research on distributional similarity has focused on the notion of *attributional similarity*, which estimates the similarity between entities (such as two drugs). However, an important component of QA involves identifying relationships between concepts.

Therefore, *relational similarity*, the estimation of the similarity between pairs of entities (such as two drug-gene *pairs*) based on the nature of the relationship between them is important. Relational similarity is estimated from the contexts in which these entity pairs occur together, and may help identify interesting relationships between biomedical concepts. However, within biomedicine scant research exists on this topic. Methods for estimation of relational similarity have seldom been evaluated, and little is known about how these methods might be leveraged for QA purposes in emerging domains.

Hypothesis and Specific Aims

The dissertation explores the utility of a scalable corpus-based approach to estimate the relational similarity between pairs of concepts extracted from Medline abstracts. *My hypothesis is that relational similarity will be of greater utility than attributional similarity as a means to identify biological relationships that may provide answers to clinical questions, (such as "which drugs INHIBIT gene x"?) in the context of rapidly evolving domains.*

In the context of the application domain of precision oncology, I evaluate this hypothesis using sets of known relationships as seeds, and attempting to generalize from them using both attributional (which drugs *are similar to* the known inhibitors of x?) and relational (which drugs *relate to* gene x *in a similar manner* to known inhibitors of x?) similarity, with the following Specific Aims:

Aim 1: Develop and implement models of attributional and relational similarity

Models of relational and attributional similarity are developed using two widely-used distributional semantics techniques: Random Indexing (M. Sahlgren, 2005) (RI) and Neural Word Embeddings (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Yih, & Zweig, 2013) (NWE).

Aim 1.1 Relational similarity.

With RI, I *explicitly* identify drug-gene pairs, and derive vector representations of these *concept pairs* from the terms that occur between them. The similarity between the resulting *pair vectors* is used to draw inference about previously unseen pairs. With NWE, I use *implicit* relational information by performing geometric operations on *concept vectors* ($\overrightarrow{Drug_{cue}} - \overrightarrow{Gene_{cue}} + \overrightarrow{Gene_{target}} \cong ?$ (Mikolov, Chen, et al., 2013). Relational similarity is estimated as the cosine metric between the vector resulting from these operations, and the NWE vector for a candidate drug.

Aim 1.2 Attributional similarity.

With both RI and NWE I use Medline abstracts as *documents* to build vector spaces, and measure the cosine similarity between *concepts*.

Aim 2: Recovery of held-out drug/gene relationships

Using a reference set of clinically-relevant drug/gene relationships developed for precision oncology, the models from SA1 are evaluated for their ability to recover held-out relationships given a set of seed examples, across a broad range of cross-validation configurations.

Aim 3: Unsupervised identification of clinically relevant drug/gene relationships

As implemented to meet Aim 2, relational similarity models require a set of expert-generated "seed" examples to serve as cues. As these examples may be unavailable at the outset of a project, in this Aim I develop and evaluate an alternative proposal in which cues are derived *without expert input*, using knowledge extracted from the biomedical literature using NLP. The attributional and relational models developed in Aim 1 are evaluated for their performance, using a reference set of clinically-relevant drug/gene relationships.

Biomedical relevance

Although the methods developed and evaluated in this dissertation should be applicable to identifying biomedically meaningful relationships in general, I have selected Precision Oncology, the use of molecular characteristics of a tumor and patient attributes, to "personalize" therapy, as an application domain on account of the pressing need for identification of clinically relevant drug/gene relationships in this domain. (Garraway, Verweij, Ballman, & others, 2013; Meric-Bernstam, Farhangfar, Mendelsohn, & Mills,

2013) To support clinical decisions, domain experts must continuously review the published literature to develop and maintain a knowledge base of cancer-related genes, and the agents that target these genes or their associated biological pathways. (Johnson et al., 2015) With both the number of genes and the relevant literature growing rapidly, manual review of the literature in search of new therapies is not scalable, and there is a pressing need for informatics technologies to help curators more rapidly retrieve and review relevant biomedical literature. (Johnson et al., 2015; Meric-Bernstam et al., 2013) The methods developed and evaluated in this project can serve as an important step toward that goal.

Guide for the reader

The remainder of this dissertation proceeds as follows. Chapter 2 provides an in-depth review of the literature on distributional semantics, relational and attributional similarity, theoretical and cognitive basis of relational similarity, question answering, and informatics needs of precision oncology. Chapter 3 describes the details of my preliminary experiments, and in particular AIMED(Fathiamini et al., 2016), an application built to retrieve drug/gene relationships from biomedical text, which elucidates some of the challenges of this task in the domain of precision oncology, and helps form a basis for the next experiments. Chapter 4 reports on the details and results of the Specific Aims 1, and 2 of the research. Chapter 5 discusses the results of the Specific Aim 3, and their significance. Chapter 6 summarizes the accomplishments, contributions of the research described in this dissertation, and future work.

Chapter 2: Literature Review

As explained in Chapter 1, the focus of this research is on a comparative analysis of a range of relational and attributional similarity techniques – components of the broader field of distributional semantics – and the application domain I have selected in which to do this evaluation is precision oncology. As part of my preliminary studies to better understand the characteristics and informatics requirements in this field, I created a Question/Answering (QA) application to help a team of curators find the answers to their questions of type "What drugs inhibit gene X?", and maintain a knowledge base of drug-gene relationships. This project helped elucidate some of the unique challenges of this task in the domain of precision oncology, and led us to realize the need for the current research. Some of the text in this chapter is borrowed from our published paper from this project. (Fathiamini et al., 2016)

To follow the natural progression of ideas that led to the conception of the current research, I will present my findings from the existing literature in the following order: First I will briefly discuss QA systems, with a focus on biomedical QA, and in particular, as it applies to emerging domains such as precision oncology. Next, I will touch on techniques of relationship extraction, and make a case of why methods of distributional semantics may be particularly valuable in this domain. Finally, I will present the recent developments in

relational and attributional similarity methods, and explain the need for further research in this domain.

The challenge of biomedical information retrieval

The biomedical literature often contains answers to clinicians' clinical and research questions, (Westbrook, Coiera, & Gosling, 2005; WESTBROOK, GOSLING, & PSYCHD, 2004) and clinicians believe that the quality of patient care could be improved by online search.(WESTBROOK et al., 2004) However, the answers to two-thirds of the questions that clinicians have about their patients are either not pursued, or pursued but not found. (Chambliss & Conley, 1996; Currie et al., 2003; Huang, Lin, & Demner-Fushman, 2006) Further analysis shows that poorly constructed queries is one of the main reasons why the right answers cannot be found. (Demner-Fushman & Lin, 2007; Gorman & Helfand, 1995) Besides, given the overwhelming size of the documents that are often returned by PubMed/MEDLINE, identifying relevant citations can be difficult, and advanced features such as Boolean combinations of MeSH terms are seldom used.(Haynes et al., 1990; Herskovic, Tanaka, Hersh, & Bernstam, 2007) Also physicians may be concerned about existence of answers, have time limitations, or have doubts about the optimal search strategy.(Ely, Osheroff, Chambliss, Ebell, & Rosenbaum, 2005; Ely et al., 2002) They spend much less time searching for an answer than would be required to find one.(Ely et al., 1999; W. R. Hersh et al., 2002) In general, the ability of the users to find answers to their clinical questions using Medline is low. (W. R. Hersh et al., 2002) QA systems have been proposed as a solution to this problem.(Athenikos & Han, 2010)

Biomedical QA systems

Traditionally, document retrieval systems (such as PubMed) return a list of documents in response to a user's query. However, this requires manual review of each document. So, QA systems that return structured knowledge (e.g., drug A targets gene B) with links to supporting documents are a desirable alternative.(Athenikos & Han, 2010; W. R. Hersh & SpringerLink (Online service), 2009; Voorhees, 2001) Given the rapid growth of online literature, it has been argued that QA capabilities are among the most critical features of future search engines.(Athenikos & Han, 2010) QA systems try to provide accurate answers to their questions by integrating Natural Language Processing (NLP), text summarization, information extraction, and statistical and knowledge-based methods.(Demner-Fushman, Chapman, & McDonald, 2009; Hirschman & Gaizauskas, 2001) Early QA systems only relied on term based methods to generate answers. However, due to the availability of vast amounts of biomedical information, and its crucial role in research and applications, there was a growing need for better QA systems that could help researchers and healthcare professionals in their search for answers to their questions. (Athenikos & Han, 2010) As such, biomedical QA systems moved beyond the surface level term based analysis, drawing on knowledge-based ontological resources.(Athenikos & Han, 2010)

Knowledge-based QA systems

A wealth of knowledge resources, including ontologies, have been developed in biomedicine over the past few decades that can be used by computers when processing

complex queries, and there is evidence that they are of value for QA. (Rinaldi, Dowdall, & Schneider, 2004; Yu & Sable, n.d.; Zweigenbaum, 2003, 2009) To provide accurate answers, most QA systems in biomedicine draw upon these curated knowledge sources (such as the Unified Medical Language System or UMLS), and leverage the reasoning capabilities that ensue to address issues such as ambiguity and synonymy, and also facilitate cross document or cross knowledge-base queries using inference.(Athenikos & Han, 2010; Lopez, Motta, Uren, & Sabou, 2007) Analysis of the TREC Genomics Track ("TREC Genomics Track," n.d.), which focused solely on biomedical content and was one of the largest challenge evaluations in biomedical QA, showed that normalization of query terms and use of the Entrez Gene thesaurus for synonym expansion, post-filtering answers, and the option to specify answer entity types (e.g., genes, proteins, diseases, etc.) were among the factors associated with higher performance. (W. Hersh, Cohen, Ruslen, & Roberts, 2007; MOLDOVAN, CA, HARABAGIU, & SURDEANU, 2003; Rekapalli, Cohen, & Hersh, 2006)

However, structured knowledge alone is not adequate to obtain state-of-the-art performance. The majority of medical QA system use a combination of knowledge based and statistical methods to find their answers.(Athenikos & Han, 2010) For example, *CQA-1.0* (Demner-Fushman & Lin, 2007) is a semantics-based medical QA system based on the PICO framework – a guideline of evidence-based medicine (EBM), stating that constructing a clinical question in terms of the four areas of Problem/Population, Intervention, Comparison, and Outcome (PICO) facilitates searching for an accurate

answer (Richardson, Wilson, Nishikawa, & Hayward, 1994). It uses a combination of statistical methods (including supervised machine learning) and knowledge-based techniques (leveraging the UMLS and MetaMap (Aronson, n.d.)) to identify relevant Medline abstracts, ranks them using a multi-step scoring system, and returns short passages as answers. *Essie* (Ide, Loane, & Demner-Fushman, 2007) is a probabilistic search engine developed at the National Library of Medicine for the ClinicalTrials.gov database, and provide a concept-based search using UMLS-derived synonymy, document relevance ranking using positional information (such as location in the document with regard to different sections) of the search phrase, and query expansion using UMLS SPECIALIST lexicon (McCray, Srinivasan, & Browne, 1994). Essie was the best performing search engine in 2003 TREC Genomics track (SNEIDERMAN, DEMNER-FUSHMAN, FISZMAN, IDE, & RINDFLESCH, 2007), and one of the best in 2006. (Ide et al., 2007) *SEM-BT* (Hristovski, Dinevski, Kastrin, & Rindflesch, 2015) is a biomedical search engine that implements QA as a search in a database of semantic relations, extracted from biomedical literature by SemRep NLP system (Rindflesch & Fiszman, 2003), a natural language processing tool developed at the National Library of Medicine. SemRep depends upon both MetaMap (Aronson, n.d.) and knowledge encoded in the UMLS.(Bodenreider, 2004) *MiPACQ* (Cairns et al., 2011) is a clinical QA systems that integrates multiple Natural Language Processing (NLP) components to achieve deep semantic understanding of medical questions and texts. MiPACQ provides query formulation, automatic question and candidate answer annotation, and machine learning (ML) based answer re-ranking. *AskHERMES* (Cao et al., 2011) is a clinical QA system that performs semantic analysis on

clinical questions and outputs question-focused extractive summaries as answers. The system indexes five types of resources: MEDLINE abstracts, PubMed Central full-text articles, eMedicine documents, clinical guidelines and Wikipedia articles. In an experiment three systems (SemRep, Essie, and CQA-1.0) were examined in combination, to determine how traditional information retrieval (Pubmed search) could be improved using knowledge-based methods in a hybrid approach to question answering. Those systems used varying degrees of semantic knowledge, and overall, combining those methods resulted in better system performance than that of individual systems.(SNEIDERMAN et al., 2007)

There are medical QA systems that do not employ a knowledge-based approach. MedQA (Lee et al., 2006) for example, uses a syntactic parser for question classification, standard IR methods for document retrieval, and syntactic and statistical techniques such as document zone detection and clustering for answer extraction. Still, the creators of this system recognize the need for a domain specific parser and the importance of capturing semantic information, and the need for UMLS concepts and semantic types to help classify questions more effectively. (Yu & Cao, 2008; Yu, Sable, & Zhu, n.d.) What these systems have in common is reliance on domain-specific knowledge resources. This dependence is likely to be a liability in emerging domains.

Medical QA in emerging domains

The application domain I have selected in which to evaluate the relative merits of attributional and relational similarity is precision oncology. This task-domain is different than those that have motivated the development of prior QA systems. Typically, medical

QA systems follow an EBM-based approach, and try to provide answers supported by extensive evidence. In rapidly evolving domains such as precision oncology, the resources found in the clinical literature are often scarce, and relation extraction systems that rely on well-established knowledge and favor precision over recall (such as SemRep) may miss valuable information (On average, SemRep provides a precision of around 77% across different experiments (Kilicoglu et al., 2008), and in one study its recall was around 55%. (Ahlers, Fiszman, Demner-Fushman, Lang, & Rindflesch, 2007)) Further, in order to provide accurate answers, medical QA systems often draw upon manually constructed ontologies and leverage semantic classes or domain specific typology of questions to provide more accurate answers, or limit the size of their result sets. However, the utility of such semantic resources is restricted to topics where the concepts and relationships have already been defined, usually based on well-established knowledge. Due to the rapid emergence of new knowledge in emerging domains, there is often a knowledge gap between the newly discovered entities and relationships, and those described in existing ontologies (in precision oncology, an example might be discovery of new drugs that are yet to be added to existing drug ontologies). Furthermore, knowledge from both the literature (including clinical and cancer biology) and other sources (such as clinical trials or pharmaceutical companies) may be relevant, which presents additional challenges for the technologies employed. For example, pharmaceutical companies do not expose their drug annotations as structured data, and the need to extract this information from web pages introduces additional complications and vulnerabilities to error.

As part of preliminary experiments to partially address this problem, we introduced the AIMED system (Fathiamini et al., 2016) (explained in Chapter 3). In this system, I showed that the knowledge-driven SemRep biomedical NLP system was only beneficial for finding established drugs, whereas with investigational agents, the performance was better when using co-occurrence counts without the use of NLP (other than for concept extraction and normalization). However, while recall improved with the use of co-occurrence, precision decreased since extracted relationships were no longer constrained. These results revealed an underexplored area between the linguistic rules and semantic constraints that systems such as SemRep impose to identify specific relationships on the one hand (thus achieving higher precision), and the unconstrained associations defined by co-occurrence (evident by higher recall) on the other. In the absence of established relationships as the underlying knowledge to constrain Boolean retrieval, the co-occurrence result sets can be overwhelmingly large. One approach to this problem involves applying relationship extraction techniques to find only those relations that are relevant to the query. A general overview of relationship extraction is discussed next.

Relationship extraction

There is a large body of research concerning relationship extraction (RE), and NLP methods that can analyze text and find the relationships of interest in biomedical domains.(Friedman, Kra, Yu, Krauthammer, & Rzhetsky, 2001; Fundel, Küffner, & Zimmer, 2006; Kotecki & Cochran, 2002; McDonald et al., 2005; Rindflesch & Fiszman, 2003) The goal of RE is to identify a relationship between a pair of entities of specific

types. The relationship can be general (like any biological relationship) or specific (such as an INHIBITS relation).(A. M. Cohen & Hersh, 2005) Biomedical RE is often considered a sentence level problem which relies on rules or ontologies that map terms to standard concept representations such as UMLS Concept Unique Identifiers (CUIs). Maintaining these representations, and subsequently, building bio-medically relevant models based on them (that need to be rebuilt for each new domain) is time consuming and requires constant human supervision and effort.(Trevor Cohen & Widdows, 2009) Due to inaccuracies in the NLP process in general, and RE from the biomedical literature in particular, many biomedical knowledge bases such as PharmGKB and DrugBank are entirely based on manual curation.(Percha & Altman, 2015) My Cancer Genome is another example of a manually curated knowledge base that provides precision oncology related resources.("My Cancer Genome, Genetically Informed Cancer Medicine," n.d.) Similarly, the Drug-Gene Interaction database (DGIdb) is a database of potentially druggable genes aggregated from multiple other resources including My Cancer Genome and other manually curated databases.(Griffith et al., 2013) Another set of techniques for sentence level RE apply machine-learning methods, avoid rules, but require annotated sentences for training. Such annotation is time consuming and human intensive.(Kim, Ohta, Tateisi, & Tsujii, 2003) Linguistic patterns (such as regular expressions) have been used for RE, either as prescribed by domain experts, or automatically by generalizing patterns from training sets and searching among sentences to find commonalities. These methods take a long time to process text and generate results, especially with large pattern sets. (A. M. Cohen & Hersh, 2005; Hakenberg et al., 2010)

Distributional semantics

In contrast to sentence-level approaches, statistical methods have been applied to find concepts that co-occur with each other more frequently than would be observed by chance. It has been argued that this corpus-based approach provides a more robust mechanism for finding relationships of interest, as it infers relationships from the overall distribution of terms across an entire corpus of text, rather than from an individual assertion.(Percha & Altman, 2015) These methods, collectively known as *distributional semantics* (Trevor Cohen & Widdows, 2009; Levy, Goldberg, Dagan, & Ramat-Gan, 2015) correspond to cognitive models of memory recall (Trevor Cohen & Widdows, 2009; Kanerva, 2010; Landauer, Foltz, & Laham, 1998; LUND & BURGESS, n.d.), and match well to human judgment of pairwise correlation between biomedical concepts. (McInnes & Pedersen, 2017; Pakhomov, Finley, McEwan, Wang, & Melton, 2016) They provide fast and robust mechanisms to find relatedness and similarity between concepts and relations (Trevor Cohen & Widdows, 2009), and have been used to identify relationships between entities. (Lin & Pantel, 2001; Riedel, Yao, Marlin, & McCallum, 2013) These models can also capture information concerning the nature of the relationships between terms, either incidentally (Mikolov, Yih, et al., 2013), or by design (Turney, 2005). An important distinction concerning the nature of the estimates derived from these models is that between attributional and relational similarity.

Attributional similarity

The majority of the research on distributional semantics has focused on *attributional* similarity – similarity between objects or their properties. (Medin, Goldstone, & Gentner, 1990) That is to say, distributional methods have been developed and evaluated for their ability to capture the similarity between conceptually-related entities. This is possible because distributional methods enable the estimation of a quantitative measure of semantic relatedness between terms from the contexts in which they occur in across a large corpus of text. Geometric approaches to this problem involve the derivation of reduced-dimensional (i.e. with dimensionality less than the number of unique contexts, or context terms in a corpus) vector representations of terms from the contexts in which they occur (Turney & Pantel, 2010), such that terms that occur in similar contexts will have similar vector representations. The distance between the resulting vectors provides a meaningful estimate of semantic similarity and relatedness. Such approaches include (but are not limited to) Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997), and the Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996), which have been used to find similarity between terms and documents with good correspondence with human performance across a range of cognitive tasks. (Landauer et al., 1998; LUND & BURGESS, n.d.) Another method, Random Indexing (M. Sahlgren, 2005) generates a reduced dimensional space and produces similar results to LSA in evaluations of the quality of term-term similarity such as synonym tests, and correspondence with free association norms (Kanerva, Kristoferson, & Holst, 2000; Magnus Sahlgren, 2006), while requiring

much less computational power.(Trevor Cohen & Widdows, 2009; M. Sahlgren, 2005) These methods have been applied to problems such as information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, n.d.), literature-based knowledge discovery (Gordon & Dumais, 1998), bilingual information extraction (Widdows et al., 2003), and relationship extraction (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). More recently, neural word embeddings (Mikolov, Chen, et al., 2013; Mikolov, Yih, et al., 2013) have become a popular method of generating such reduced-dimensional representations, with improvements over prior distributional methods in some evaluations (Baroni, Dinu, & Kruszewski, 2014) (although some of these improvements have been shown to be contingent upon optimal configuration of model hyper-parameters in subsequent experiments (Levy et al., 2015)).

Relational similarity

Relational similarity, on the other hand, involves similarity between any two given pairs of concepts – if A's relationship to B is similar to C's relationship to D, then A::B is in relational similarity to C::D. Theories of analogy seem to agree that relational similarity is a fundamental component of analogical reasoning. (GENTNER, 1988; Medin et al., 1990; Medin, Goldstone, & Gentner, 1993) According to those theories, similarity requires a point of reference – one must specify the aspect from which two things are similar (e.g. drugs can be similar based on their clinical effect, chemical composition, etc.) – and in the case of relational similarity, the relational commonalities provide the relevant aspect of similarity (as in *"Drug A and B are similar based on their relationship to gene C"*).

(Goodman, 1972; HOLYOAK & THAGARD, 1989; Medin et al., 1993) In the section that follows, I review some of the recent work in distributional semantics that has attempted to estimate structural similarity of this sort from text directly.

In seminal work in this area, Turney and Littman created a Vector Space Model (VSM) for calculating relational similarity.(Turney & Littman, 2005) Sixty-four "joining words" (such as "for", "of", "to", etc.) were used to create patterns of both "A *join* B" and "B *join* A" (such as "A of B", "B of A", etc.). Then, they characterized the relationship between two words (A and B) by counting the number of times they appeared in those patterns across the corpus, which yielded a vector of 128 numbers for each A::B relationship of interest. The relational similarity between any two given pairs of words was then represented by the cosine similarity between their corresponding vectors.(Turney & Littman, 2005) This work was then extended by Turney to develop Latent Relational Analysis (LRA) [47], a technique for measuring relational similarity that adapts the VSM model in three ways: 1) patterns are extracted from the corpus dynamically by finding exemplar phrases that involve the pair of interest, 2) a thesaurus is used to extend the search space by including words that are similar to the query terms (the pair), and 3) in a manner reminiscent of LSA, Singular Value Decomposition (SVD) is used for dimension reduction of the resulting pair-by-pattern matrix.(Turney, 2005) As such, LRA may be inconvenient to implement, particularly when pairs of interest change frequently and the text corpus is large, and may scale poorly to large sets of concept pairs on account of the need to apply the SVD. For example, in one 2005 experiment it took LRA nine days to return results for

374 analogy test questions, running on a matrix of 8,000 columns and 17,232 rows (48 pairs per question, with some omissions). Although the software was not optimized for speed (Turney, 2005), and the decomposition would no doubt run faster on contemporary hardware, decomposing a matrix with as many rows as there are concept pairs of interest is a computationally inconvenient feature of this model.

Recent work in the general domain has attempted to estimate relational similarity from term (rather than pair) vector representations directly, finding that word vectors derived from a scalable neural network model can implicitly capture information of this sort. Specifically, Mikolov and his colleagues developed two neural network architectures, continuous bag of words (CBOW) (which learns to predict a word based on the words that surround it), and the continuous skip-gram model (which learns to predict context words based on an observed word). These "word embedding" architectures were used to train word representations from large corpora (billions of words), and the resulting word vectors were shown to capture relationships between words, which could be recovered with simple geometric operations. For example, using the resulting vector representations, $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Germany} \cong \overrightarrow{Berlin}$.(Mikolov, Chen, et al., 2013; Mikolov, Yih, et al., 2013) Training of these architectures occurs through a process of online learning (Shalev-Shwartz, 2011), in which each training context (a "sliding window" of words surrounding each observed word) is considered independently (though global term frequency statistics are used to inform subsampling strategies). This permits parallel implementation of the training process, enhancing scalability. Alternatively, it has been shown that it is possible

to capture such relational information without training a predictive model (such as a neural network) on an example-by-example basis. Specifically, Pennington et al. introduce Global Vectors (GloVe (Pennington, Socher, & Manning, 2014)), a model for the unsupervised learning of word representations that utilizes global distributional statistics directly, while nonetheless capturing similar structural information to online neural-probabilistic methods. In some experiments, GloVe performed better than comparable neural network approaches in evaluations on pairwise analogies (Pennington et al., 2014), however these advantages were not replicated in subsequent experiments in which hyperparameters and training corpora were consistent across models (Levy et al., 2015).

Some research on relational similarity exists in the biomedical domain. Predication-based Semantic Indexing (PSI) is a variant of Random Indexing that explicitly encodes relationships between concepts from a collection of semantic predications (such as those extracted by SemRep, for example *docetaxel* STIMULATES *akt*) into distributed vector representations of these concepts.(Trevor Cohen, Schvaneveldt, & Rindflesch, 2009; Trevor Cohen, Widdows, Schvaneveldt, Davies, & Rindflesch, 2012; Widdows & Cohen, 2015) Across several experiments (see for example (T. Cohen et al., 2014; Trevor Cohen et al., 2012; Shang, Xu, Rindflesch, & Cohen, 2014), and for a review in (Widdows & Cohen, 2015)), PSI was applied to infer previously unseen relationships by using relational similarity, including both harmful and potentially therapeutic drug/effect relationships.(Trevor Cohen, Widdows, Schvaneveldt, & Rindflesch, 2011) Embedding of Semantic Predications (ESP) is a neural-probabilistic alternative to PSI that has shown

advantages in predictive modeling experiments using estimates of relational similarity. (Trevor Cohen & Widdows, 2017) Both PSI and ESP use relations extracted by SemRep rather than free text, and thus represent a different class of methods to those under consideration in the current work.

Of particular relevance to the current work, Percha and Altman developed a method that uses grammatical dependency paths in the sentences that contain a pair of concepts as contextual features. (Percha & Altman, 2015) An unsupervised clustering technique called Ensemble Biclustering for Classification (EBC) is then applied to the resulting pair-by-path matrix, such that drug-gene pairs are represented by their frequencies of co-clustering with every other pair across large numbers of stochastically-initialized clustering processes. As drug/gene pairs linked by similar dependency paths will cluster together, EBC leverages relational similarity drawn from distributional statistics. Using a seed set of ten drug-gene relationships, EBC was shown to successfully detect similar relations from a large corpus of Medline abstracts.(Percha & Altman, 2015) The relations identified in this process were recently made publicly available (Percha, Altman, & Wren, 2018a). Because operates in a largely unsupervised manner, EBC is not readily adaptable to the cue/response paradigm I employ in the current evaluation, which is limited to methods that do not require parsing to reveal grammatical dependencies.

Summary of research on relational similarity

The techniques discussed above have been mostly applied in the general domain, resulting in the development of online techniques, such as random projections and neural word

embeddings, that can be used to create relational similarity models without requiring computationally demanding techniques of dimension reduction. Given the size of the pharmacogenomics search space, this is an important consideration. PSI and ESP are similarly scalable, but explicitly encode relations extracted by SemRep, and therefore are in a different methodological category to those methods attempting to infer relational information from free text directly. Dealing with text directly is a desirable alternative in emerging domains, on account of the time lag in the incorporation of emerging drugs into the knowledge sources upon which NLP systems such as SemRep depend, and the fact that SemRep's optimization for precision over recall is not ideal for concepts that appear in a small number of citations only. EBC searches for the relational similarity between drug-gene pairs by applying distributional techniques across Medline abstracts, but uses only one type of linking relationship (dependency paths), and has not been evaluated against an attributional counterpart. While it seems intuitive that *relational* similarity would be better suited to recognition of biomedical *relationships* than attributional similarity, this hypothesis has not been tested.

Overall, there is an opportunity for further research to identify techniques based on relational similarity to identify meaningful drug/gene relationships in emerging biomedical domains. The current research explores the application of relational and attributional similarity techniques in precision oncology, as an exemplar of an emerging biomedical domain, focusing specifically on drug-gene relationship extraction from Medline abstracts.

Chapter 3: Preliminary experiments – Automatic Identification of Molecular Effects of Drugs (AIMED)

My preliminary work examines the utility of relatively constrained semantic relationships versus relatively unconstrained co-occurrence statistics. The results of this research revealed an underexplored area between these two ends of the relationship extraction spectrum, and motivated the development of a hypothesis that forms the theoretical basis for this dissertation. The evaluation explained in this chapter was conducted in the context of a QA system I developed to find relevant drug-gene relationships in the context of precision oncology, which provides the practical motivation for the specific aims of this dissertation. Some sections in this chapter are borrowed from my previously published paper (Fathiamini et al., 2016).

Precision oncology

Precision oncology, or personalized cancer therapy, involves the use of molecular characteristics of a tumor and patient attributes, to "personalize" therapy with the goal of more effective and less toxic cancer treatment.(Garraway et al., 2013; Meric-Bernstam et al., 2013) Therapy can be personalized using different aspects, including a specific patient's exposure history, preferences and clinical features. However, genomic profiling is emerging as a popular personalized option that is affordable, increasingly available to

cancer patients, and can help select “genomically-informed” targeted therapy options, and oncologists can prescribe treatment targeted to specific molecular aberrations found in a patient’s tumor.

To support clinical decisions, domain experts must continuously review the published literature to develop and maintain a knowledge base of cancer-related genes, and the agents that target these genes or their associated biological pathways.(Johnson et al., 2015) Personalizedcancertherapy.org is one such knowledge base that can serve as a reference for clinicians.(Johnson et al., 2015) Existing technologies that extract knowledge from the biomedical literature are generally designed for stable domains where the state of knowledge evolves relatively slowly. For example, one analysis found that 90% of clinical practice guidelines were still valid at 3.6 years. (Shekelle et al., 2001) In contrast, precision oncology evolves much more rapidly. As information concerning newer agents is relatively scarce, established relation extraction systems that rely on established knowledge resources and favor precision over recall (such as SemRep (Rindflesch & Fiszman, 2003), an NLP tool developed at the National Library of Medicine) may miss valuable information. Further, many targeted therapies are investigational and are currently available primarily via clinical trials. Thus, there is an urgent need to develop informatics technologies to help curate pertinent clinical information.

To this end, I developed a system for the Automated Identification of Molecular Effects of Drugs (AIMED), which leverages semantic information extracted by the SemRep and MetaMap (Aronson, n.d.) NLP systems, but augments this using task-specific filtering of

results and drug-gene co-occurrence data to extract clinically relevant pharmacogenomic relationships from the biomedical literature.

Materials

In this section I introduce the tools and materials that I have used to create AIMED.

SemRep_UTH, a modified version of SemRep

I designed and implemented a semantic QA system based on a large collection of predications that is publicly available in SemMedDB,(Kilicoglu, Shin, Fiszman, Rosemblat, & Rindflesch, 2012) which is generated by SemRep processing of Medline abstracts. Semantic predications in SemMedDB are organized as *Subject-Predicate-Object* triples, with subjects and objects being UMLS concepts, and predicates coming from UMLS Semantic Network. The Semantic Network defines allowable relationship types between any two concepts, based on their semantic type. (At, 1989) In early experiments I realized that many of the drugs that were relevant to precision oncology were underrepresented in this database. A search for some of these drugs (such as *AZD2014*) in online versions of interactive SemRep and MetaMap (accessible at ("Interactive SemRep," n.d.) and ("Interactive MetaMap," n.d.), respectively) revealed that SemRep by default uses a rather old version of UMLS (2006), and it "suppresses" some of the short forms of drug names, many of which relevant to my work (for example, AZD2014 was only recognizable by MetaMap, and so SemRep, in its full form as "*mTOR Kinase Inhibitor AZD2014*"). To address this problem, I updated SemRep's data files to the latest version of the UMLS at the time of this experiment, 2013AB. To do this, I installed UMLS locally, and modified

the MRCONSO.RRF file. First I identified terms with both SAB (Abbreviated Source Name) value equal to 'NCI' (National Cancer Institute) and with TTY (Term Type) value equal to 'CCN' (Chemical Code Name) (henceforth: NCI/CCN). For NCI/CCN terms, I changed the 'SUPPRESS' value from 'Y' to 'N'. This caused those terms to become 'active' and therefore be useable by MetaMap and SemRep. Also, for NCI/CCN terms I changed the value of *TS* (Term Status) to 'P' ('Preferred') where they were 'S' ('Synonym'), so that they would all be chosen by MetaMap and SemRep when encountered. I then used MetaMap Data File Builder ("MetaMap Data File Builder," n.d.) to compile UMLS files and make them available to SemRep. As an additional step, I also removed –D flag from SemRep to identify more concepts. Leaving –D in place would block 'dysonyms', certain UMLS synonyms that are considered harmful. This version of SemRep (henceforth: SemRep_UTH) was used throughout this project for extraction of semantic predications and to normalize drug and gene names (explained below).

SemMedDB_UTH an enhanced repository of semantic predications

I used 23,537,576 PubMed abstracts downloaded in August of 2014 (henceforth: PMAbstracts), as my knowledge source. I processed PMAbstracts (mentioned above) using SemRep_UTH, and created SemMedDB_UTH (hosted by the NLM[1]) ("SemMedDB_UTH Database Outline," n.d.). This database is similar to the original SemMedDB in that it follows the same database schema, but contains more predications (especially drugs and

[1] https://skr3.nlm.nih.gov/SemMedDB/index_uth.html

genes that are important for the purposes of this project), as a result of updated data files. I also added a new table, ENTITY, which contained all the concepts recognized by SemRep (not just the ones used to create predications). I used the information from the ENTITY table to create co-occurrence relationships between drug-gene concept pairs at the sentence, and document level. In a similar fashion, I used the 'summary' and 'full description' sections from 183,260 trials downloaded from ClinicaltTrials.gov (https://clinicaltrials.gov/) in January of 2015 (henceforth: CTDescs), and added them to my data source. I used an original version of SemMedDB, v. 23 (Kilicoglu et al., 2012; Rindflesch et al., 2011; "SemMedDB Info," n.d.) as the baseline to which I compared my results.

Drug-Gene relations reference set

To evaluate the results of my queries, I used as the reference set the gene-drug knowledge base (henceforth: Gene Sheets) (accessible at http://personalizedcancertherapy.org, with permission) provided and maintained by 12 cancer biologists and clinicians at the Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy (IPCT) Precision Oncology Decision Support team at the MD Anderson Cancer Center. Each gene sheet contained a list of drugs that are relevant for tumors with alterations in that gene. The Gene Sheets describe genetic pathways known to be involved in certain cancer types. Each pathway includes a main gene and a list of downstream genes that are thought to be of interest as alternative therapeutic targets in the event the main gene cannot be directly targeted. I used the gene *PIK3CA* for my formative evaluation (preliminary experiment

below), and 17 other genes, *ABL1, AKT1, ALK, BRAF, CDK4, CDK6, EGFR, ERBB2, FGFR1, FGFR2, FLT3, KDR, KIT, PDGFRA, RET, ROS1, SMO* for the summative evaluation that followed. All the gene and drug names were normalized to UMLS concept unique identifiers (CUIs), or Enterz Gene IDs (for genes only) using SemRep_UTH.

The Semantic Query

The semantic query was formulated to represent a query for *"drugs that target genes [of interest]"*. I mapped the verb 'target' to different relationship types (predicates) at different stages of my project to use in the query. In the preliminary experiment (see below), I chose *INHIBITS* and *INTERACTS_WITH* to use in the query. The predicate 'INHIBITS' was chosen because all the genes in the development and test set were oncogenes (rather than tumor suppressor genes), and the goal was to find inhibitors of these genes. The predicate 'INTERACTS_WITH' was chosen by examining the existing predications from SemMedDB_UTH, observing it tended to represent relationships pertinent to targeted therapy. In the final stage of the project (evaluation phase, see below) I added *COEXISTS_WITH* based on the insight gained from tests on a "development set" (see below) used to find the optimal set of system parameters. The query would then involve finding drugs (output) that were in a certain relationship (predicate) with a known list of genes (input). The predicates were all bi-directional (with the exception of INHIBITS), so I treated them as such, i.e. I looked for relationships in both directions. I looked for any drug that targeted the main gene, any gene downstream, or any of their synonyms. Downstream genes and synonyms are provided as part of the Gene Sheets. For some of

the gene synonyms, no normalized form was found by SemRep, and they were excluded from the analysis. I limited the query to certain semantic types. For genes, I used gngm, aapp, enzy, and for drugs I used orch and phsu in the preliminary experiment, adding antb, clnd, horm, imft, nnon, opco, aapp for the final stage of the project. The choice of semantic types was made by examining the list of available semantic types in the UMLS ("Semantic Types and Groups," n.d.), and choosing the ones relevant to precision oncology. The choice was eventually verified based on the results from the result from the "development phase" (See section on "Parameter selection" below). Table 1 shows these semantic types. Figure 2 shows an overview of the system.

Table 1. *Semantic types used to create co-occurrence data*

Semantic Type	**Description**	**Representing**
aapp	Amino Acid, Peptide, or Protein	Drugs and genes
antb	Antibiotic	Drugs
clnd	Clinical Drug	Drugs
enzy	Enzyme	Genes
gngm	Gene or Genome	Genes
horm	Hormone	Drugs
imft	Immunologic Factor	Drugs
nnon	Nucleic Acid, Nucleoside, or Nucleotide	Drugs
opco	Organophosphorus Compound	Drugs
orch	Organic Chemical	Drugs
phsu	Pharmacologic Substance	Drugs

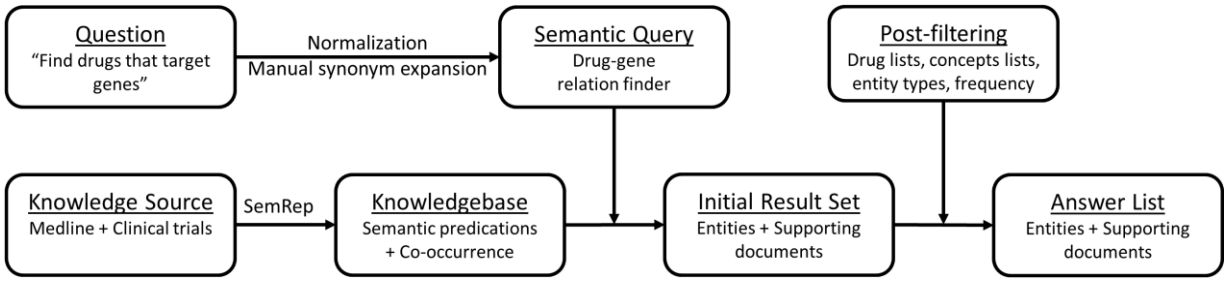

Figure 2. High level summary of the AIMED system built for finding drugs that target genes of interest.

Drug filters

The goal of this project was to find clinically-available drugs (i.e., that could be used to treat patients). Therefore, I only retained results that were either a drug in clinical trials (CT filter) or an FDA-approved drug (FDA filter). I downloaded the list of drugs from FDA (http://www.fda.gov/) and normalized them using SemRep_UTH. Also, I processed the list of drugs that were mentioned in any ClinicalTrial.gov records and normalized them using a similar method. Furthermore, drugs available via clinical trials were associated with the trial phase (i.e., phase 1, 1/2, 2, 3, with phase 1/2 involving both phase 1 and 2).("The FDA's Drug Review Process: Ensuring Drugs Are Safe and Effective," n.d.) For any given drug from clinical trials, the highest phase that could be identified was used. FDA-approved drugs were assigned phase 4. Using phase information also allowed me to limit the data source for the query evaluation. To calculate precision and recall, each drug would be considered within its phase category only. For example, to evaluate the query performance for phase 3, drugs from other phases would be eliminated from the result set, and then the

performance would be calculated against the same phase drugs from the reference set. These constraints were motivated by the assumption that the optimal strategy to identify drugs in each phase would depend on the number of drugs in this phase and the amount of published literature available concerning these drugs. I also used the information from the NCI Thesaurus (NCI filter), extracted from UMLS 2013AB, to only keep known pharmacologic substances in a systematic fashion. With this filter, I only retained drugs that were mentioned under the Pharmacologic Substance branch of the NCI thesaurus, as they appeared in the UMLS.

Preliminary experiment: Comparing SemMedDB to SemMedDB_UTH

Objective

To see whether SemMedDB_UTH has any advantage over the standard version of SemMedDB in finding drug-gene relationships

Methods

For this part of the experiment I chose one gene from the gene sheets (PIK3CA) which included two other downstream genes (AKT, MTOR), with seven drugs that would target the genes. PIK3CA was chosen as the starting point for the project as it was a current focus of IPCT discussion, and a substantial amount of related literature was already available. I ran the semantic query for this gene sheet on both databases and compared the results.

Results

In total, the number of retrieved drugs were 74 and 35, for SemMedDB_UTH and SemMedDB, respectively. SemMedDB_UTH showed a substantial advantage over SemMedDB in finding drug-gene pairs for PIK3CA-related genes. (Figure 3), increasing precision by two orders of magnitude, and identifying the remaining 70% of the reference standard drugs that were not identified using semantic queries to the original SemMedDB.

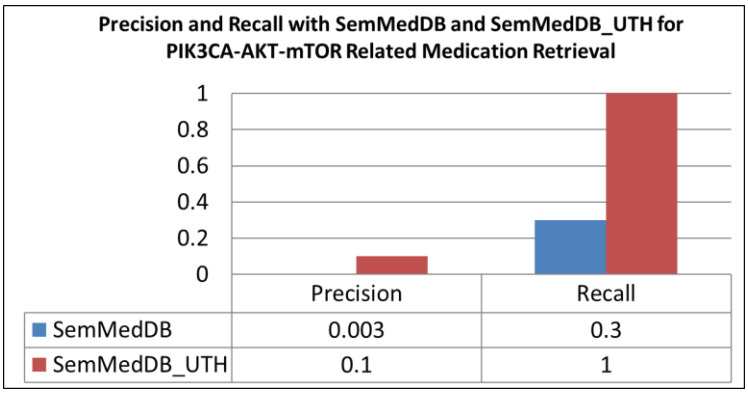

Figure 3. Results of the preliminary experiment

Discussion

SemMedDB_UTH showed a clear advantage over the standard version of SemMedDB in this experiment. The main difference between the two databases were in the underlying ontology that had been used to create them. These findings underline the importance of keeping knowledge sources up to date in this rapidly changing domain. These results also

indicated that the changes I introduced to SemRep were in fact effective, and encouraged me to continue my experiments.

Optimization of system parameters

Objective

During the development phase of the system, my goal was to find the best strategy for utilizing semantic predications and co-occurrence statistics for the task of drug-gene relationship extraction in precision oncology. Since the drugs of interest were at different development phases, one goal was to find the best set of parameters and constraints that would maximize query performance for each phase. I chose four genes as the "development set" (see next section) to test the effect of different system parameters on the query performance. The results of this development process informed the choice of parameters for the evaluation phase.

Development Set

The development set consisted of four Gene Sheets (*PIK3CA, NRAS, KRAS, MET*) chosen because they were among the first Gene Sheets developed by the IPCT, and consequently were available for development purposes while the remainder of the reference set was constructed. The development set also included the downstream genes in their respective cancer-related pathways and their known synonyms, as specified in each respective Gene Sheet. I used these four genes and their related drugs to find the best set of query parameters and constraints that would maximize the performance.

Parameter selection

Table 2 shows a summary of the query parameters and constraints used with the development set, as well as the options available for each.

Table 2. *Parameters of the system, as applied to query and the answers.*

Parameter Name	Description	Options
Semantic relationship	Type of relationship between drug and gene required for retrieval.	Predications, sentence level co-occurrence, document level co-occurrence
Food and Drug Administration (FDA) filter	Accept drugs that appear on a list of FDA approved drugs. The list was obtained from fda.gov and normalized using SemRep_UTH.	Yes/No
Clinical Trials (CT) filter	Accept drugs found in the "intervention" field from clinicaltrials.org, normalized using SemRep_UTH.	Yes/No
Phase filter	Accept drugs either passing the FDA filter (marketed) or extracted from clinicaltrials.org (CT filter) for trials with a phase of at most *x* (Phases 1-3).	Marketed, or Phase 1,1/2, 2, 3
National Cancer Institute (NCI) thesaurus filter	Return drugs that appear in the Pharmacological Substance branch of the NCI thesaurus hierarchy.	Yes/No
Frequency filter	Minimum number of extracted relationships (predication or co-occurrence) required before the drug is returned.	One to many (e.g., 5)
Predication filter	For predications, retrieve only drugs that occur in relationships with the target gene of predicate type *x*.	INHIBITS, INTERACTS_WITH, COEXISTS_WITH
Semantic type filter for co-occurrence	Semantic types of drugs to retain.	aapp, antb, clnd, horm, imft, nnon, opco, orch, phsu
Semantic type filter for predications	Semantic types to use with predication-based queries	phsu, orch

As discussed previously, the drugs that targeted the four genes in this experiment were categorized based on their development phases (i.e., clinical trial phase 1, 1/2, 2, 3 and 4 (FDA-approved)). The phase information was validated using the latest information from ClinicalTrials.gov. I considered precision, recall, F1 and F2 measures as my evaluation metrics. However, published information about potentially useful drugs may be scarce and the annotators expressed a preference for a system that would identify any potentially useful drug. Thus, recall was more important than precision, and so, I used the F2 measure (a variant of the F measure that emphasizes recall) as the single measure of choice to determine the best set of parameters within each drug phase category. The F2 is calculated as:

$$F2 = \frac{(1 + 2^2) * Precision * Recall}{(2^2 * Precision) + Recall}$$

Exploration of the space of parameters in *Table 2* in an effort to optimize the F2 metric yielded the following parameter choices: The optimal data source for marketed (phase 4) and phase 3 drugs was the semantic predications alone. This is not unexpected, as one would anticipate the availability of more published literature for SemRep processing in drugs that have advanced beyond the initial clinical trial stages. For these phases, I included results of semantic types pharmaceutical substance (phsu) and organic chemical (orch), retaining results for which at least 5 predication instances were found. For phases 2 and 1/2 I also included sentence level co-occurrence, and for phase 1 I used both predications and document level co-occurrence (with co-occurrence based on the identification of

concepts by MetaMap). Note that the set of relationships retrieved on the basis of document level co-occurrence subsumes those retrieved using semantic predications, as document level co-occurrence is a prerequisite to extraction of a semantic predication. The FDA filter for marketed drugs only, CT filter for other drug phases (3, 2, 1 /2, 1), and NCIT filter for all phases were also applied. The same set of parameters was used for Experiment 3. *Table 3* shows a summary of the final set of query parameters and constraints. *Table 4* shows the actual results of the query in the development phase that informed the choice of parameters.

Table 3. *Optimal system parameters and constraints determined in the development phase.*

Drug Phase	**Source**	**Frequency**	**Predicates**	**Drug Semantic Type**
Marketed	Predications	>4	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch
3	Predications	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	-
2	CoOcc Sen	-	-	complete list
1 / 2	CoOcc Sen	-	-	complete list
1	CoOcc Doc	-	-	complete list

Note: The same configuration was used for evaluation phase. Four Gene Sheets, and 115 related drugs were included in this experiment, and SemMedDB_UTH was used as the source of semantic predications. FDA/CT, and NCI filters were applied to all phases. CoOcc Sen: Sentence level co-occurrence, CoOcc Doc: Document level co-occurrence

Table 4. *Query results with optimal parameters for the development set.*

Drug Phase	**Documents**	**Drugs**	**Recall**	**Precision**	**F1**	**F2**
Marketed	624	50	0.86	0.12	0.21	0.39
3	242	42	0.79	0.26	0.39	0.56
2	1,466	125	0.69	0.18	0.29	0.44
1 / 2	993	25	0.45	0.20	0.28	0.36
1	544	99	0.39	0.20	0.26	0.33
All phases	**3,869**	**341**	**0.56**	**0.19**	**0.28**	**0.4**

Note: These results informed the choice of parameters in this experiment. Four Gene Sheets, and 115 related drugs were included, and SemMedDB_UTH was used as the source of semantic predications.
Documents: Number of documents returned by the query. Drugs: Number of drugs returned by the query.

Evaluate system parameters for precision oncology QA

Objective

To apply the set of parameters determined in the development phase, on a set of 17 Gene Sheets as the "evaluation set".

Methods and results

I used a set of 17 genes (*ABL1, AKT1, ALK, BRAF, CDK4, CDK6, EGFR, ERBB2, FGFR1, FGFR2, FLT3, KDR, KIT, PDGFRA, RET, ROS1, SMO*) as my evaluation set and processed them using the optimal parameters determined during system development. The gene sheets used in the development phase to identify these parameters were excluded from this evaluation. To establish a baseline, the query was also run on the standard version of SemMedDB using the same set of parameters. I found three- to four-fold improvements in

recall, precision, F1 and F2 (0.39, 0.21, 0.27, 0.33 with SemMedDB_UTH over the standard version of SemMedDB at 0.12, 0.05, 0.07, 0.09, respectively) (*Table 5*).

Table 5. *Results of the query to find drugs from the evaluation set.*

DB	Drug Phase	FDA/CT, NCI	Source	Freq.	Predicates	Drug ST	Doc.	Drug	Recall	Prec.	F1	F2
SemMedDB_UTH	Marketed	Yes	Pred.	>4	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch	2,251	80	0.69	0.3	0.42	0.55
	3	Yes	Pred.	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	-	299	61	0.35	0.3	0.32	0.34
	2	Yes	CoOccSen	-	-	complete list	4,723	205	0.5	0.17	0.25	0.36
	1 / 2	Yes	CoOccSen	-	-	complete list	3,875	40	0.29	0.18	0.22	0.26
	1	Yes	CoOccDoc	-	-	complete list	1,609	129	0.25	0.19	0.22	0.24
	All Phases						**12,757**	**515**	**0.39**	**0.21**	**0.27**	**0.33**
SemMedDB	Marketed	No	Pred.	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch	1,730	661	0.46	0.02	0.04	0.09
	3	No	Pred.	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch	1,730	661	0.17	0.01	0.02	0.04
	2	No	Pred.	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch	1,730	661	0.1	0.01	0.02	0.04
	1 / 2	No	Pred.	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch	1,730	661	0.04	0.002	0.004	0.01
	1	No	Pred.	-	INHIBITS, INTERACTS_WITH, COEXISTS_WITH	phsu, orch	1,730	661	0.01	0.002	0.003	0.01
	All Phases						**1,730**	**661**	**0.12**	**0.05**	**0.07**	**0.09**

Note: Choice of parameters from each drug phase was determined in the development phase. In total 17 Gene Sheets with 276 related drugs were used as the reference set.
Columns: DB: database used to run the queries; FDA/CT, NCI: filters used to refine the results; Freq.: frequency filter; Drug ST: drug semantic types; Doc: number of documents returned; Drug: number of concepts returned by the query; Prec: precision; F1: harmonic mean; F2: a variant of F1, emphasizing recall.

Error analysis

My error analysis focused on false negative results, as annotators had expressed a preference for a system with high recall. Of all false negative results (n=168), 19% were not found in the original knowledge sources (PMAbstracts, CTDescs). SemRep did not identify a CUI for 24%, suggesting that they did not appear in the UMLS data files used to extract concepts. Drug filters (CT/FDA, NCI) were responsible for 30% of the false negative drugs. Those drugs were either absent from the source vocabularies, or their manually designated phases were different from those specified in the filter (e.g., drug that was in phase 1 trials at the time that the reference set was created, was in phase 2 trials at the time of evaluation). Since all queries were phase based, the phase specified for the drug in the reference set had to match the one specified in the CT/FDA filter, or the drug would either be found but not matched against the reference set (wrongly marked as false positive instead of true positive), or eliminated altogether (false negative); 23% of the missing drugs would have been found if I had used a less restrictive approach, i.e., sentence level co-occurrence instead of predications (for phases 3 and 4), and document rather than sentence level co-occurrence. Finally, 4% of marketed drugs were excluded by either the frequency or semantic-type filter.

Manual evaluation

To test the hypothesis that some ostensibly false positive results were actually relevant, three domain experts from the IPCT scientific team each reviewed 50 retrieved drugs. For

each drug, experts were provided with a normalized concept name, targeted gene, a random selection of up to ten source excerpts of one or more sentences, and a link to the source document for each excerpt. To facilitate evaluation, drug and gene names were highlighted. For document level co-occurrence results, all sentences from the original document that contained the terms in question were provided. Drugs were picked in a stratified random manner from a pool of 515 retrieved drugs, equally distributed across the five phase categories (i.e., 1, 1/2, 2, 3 and marketed). Each evaluator had 40 unique drugs, and 10 drugs in common with the other evaluators to assess inter-observer agreement. Thus, a total of 135 drugs were evaluated. Each evaluator assigned a score of 1 through 3 to each source excerpt (*Table 6*).

Of the 135 drugs that were reviewed, 35 (26%) were found to receive score 3, 82 (61%) received score 2 and 18 (13%) received score 1. Inter-observer agreement was 100% (reviewers 1 and 2), 100% (reviewers 2 and 3) and 60% (reviewers 1 and 3). The drugs used to assess inter-rater agreement were different for each reviewer pair. *Table 7* shows a summary of the distribution of drugs among the reviewers.

Most of the manually reviewed results were in the score 2 group, which meant that they were relevant for review, but the level of evidence did not merit inclusion in the reference set (Gene Sheets). The score 2 group was retrospectively divided into three subcategories (high relevance – useful to communicate to clinicians but not recommended as therapy,

Table 6. *The scoring system that evaluator used to score the drug lists*

Score	Description
3	Evidence exists to add to reference set (Gene Sheets). <u>Criteria:</u> Either: • Drug directly targets and inhibits the gene *OR* • Drug indirectly targets the gene by inhibiting downstream pathway members *AND* there is evidence that alterations in the gene sensitize cells to drugs inhibiting the indirect target
2	Gene name or its alias is mentioned with the drug or its synonym, but evidence is not sufficient to add to reference set. <u>Categories:</u> **High relevance** • Indirectly targets the gene but there is no level of evidence for its use in tumors with alterations in the gene. • Partial response • Associated with resistance • Effective only in combination **Low relevance** • Mutation negative (Patients negative for mutations in a gene were treated with a drug) • Opposite association (text suggests that the gene target effects the drug, not the other way around) • Discussing an isoform or artificial version of the gene • Derivative of the drug is being discussed (not actual drug indicated in evaluation) • Association unclear • Drug targets molecule upstream of original target (not likely to be effective) • No effect **No relevance** • Not a drug/not used as a drug • No relationship/Effect untested • Drug is used as a carcinogen/ would never be used to treat cancer • Opposite effect (The drug results in increased activity of the target gene) **Not classified**
1	No mention of the drug and/or gene or its alias

Table 7. *The distribution of drugs among reviewers.*

Drug Count (Drug Number)	**Reviewer**	**Agreement**	**Details**
40 (1-40)	1		
40 (41-80)	2		
40 (81-120)	3		
5 (121-125)	1 & 2	5/5 = 100%	Both evaluator gave score 2 to all the 5 drugs.
5 (126-130)	1 & 3	3/5 = 60%	Both evaluators gave 3 of the drugs score 2. Evaluator 1 gave one drug score 2 where evaluator 2 gave it score 3. Evaluator 1 gave another drug score 3, where evaluator 2 gave it score 2.
5 (131-135)	2 & 3	5/5 = 100%	Both evaluators gave score 3 to 3 of the drugs, and score 2 to the other 2.

low relevance, and no relevance), based on curator feedback. Of note, *approximately 26% of the ostensibly false positive results were in fact relevant for inclusion in the gene sheet*. If this finding were consistent across the entire evaluation set, the re-estimated precision and recall would be 0.29 and 0.55, respectively (versus current 0.21 and 0.39, respectively). However, I cannot exclude the possibility that there are other relevant drugs that were neither retrieved by the system, nor recognized as such by our team of curators. In this case, recall may be overestimated.

Discussion

At first glance the recall, precision and F2 achieved by AIMED in the evaluation phase are relatively modest. However, manual review of ostensibly false positive results showed that 26% were actually true positives and an additional 61% were appropriate for review, but there was insufficient evidence to include these in the reference knowledge base. On the one hand this finding shows that the process of maintaining such knowledge bases (which is mostly done manually (Griffith et al., 2013; "My Cancer Genome, Genetically Informed Cancer Medicine," n.d.; Percha & Altman, 2015)) can benefit from automated systems. On the other hand, it is an indication of how this field is constantly evolving (exemplified by the progression of drugs through the development phases during the course of this work) and no "gold standard" is likely to be complete, or remain complete for long. The performance of a knowledge-based system depends on the accuracy and breadth of the source knowledge.(Basili, Hansen, Paggio, Pazienza, & Zanzotto, n.d.; Hristovski et al., 2015; Lopez et al., 2007) This is consistent with my findings, as I showed that default predications from the original SemMedDB were only modestly useful in finding emerging medications. Their utility was greatly enhanced by updating SemRep's source vocabulary, and adding predications from other knowledge sources (clinical trials). Further, we enriched the underlying ontology by modifying the data files that SemRep was using to include suppressed drug names from the NCI thesaurus. Although that technique helped with some drug categories, for drugs from lower development phases we had to further relax constraints by including co-occurrence data. Which raises the question of whether

one could just use co-occurrence, instead of any NLP-derived relationship (i.e. predications from SemMedDB_UTH in this case) to find the drugs of interest. My next experiment examines this possibility.

Comparing co-occurrence data with predications

Objective

To evaluate the utility of sentence level co-occurrence data for the task of finding drug-gene relationships of interest. In the previous experiment, I showed that combining predications with co-occurrence data can be beneficial, and the utility of each method depends on how far advanced the drug is in its development phases. Drug-gene pairs that are found in predications are always a subset of sentence level co-occurrence data, and since SemRep favors precision over recall. So, a logical assumption might be that using all drug/gene co-occurrence data, irrespective of whether a predication was identified or not, would result in better recall. In this experiment, I evaluate this hypothesis.

Methods and results

For this experiment, I designed two sets of queries. In the first set, only predications were used for all the phases, and to maximize the recall for predications, no frequency or predicate filters were applied. In the second set, only sentence level co-occurrence was used across all the phases. The results are presented in *Table 8*. Overall, the recall is 0.29 and 0.44 with predications and co-occurrence respectively, and the precision is 0.13 and 0.08 respectively.

Table 8. *Comparing predications with co-occurrence*.

DB	Drug Phase	FDA/CT, NCI	Source	Doc	Drug	True Positive	False Negative	Recall	Precision
Predications	Marketed	Yes	Predications	5046	440	30	5	0.86	0.07
	3	Yes	Predications	389	64	18	34	0.35	0.28
	2	Yes	Predications	467	91	23	47	0.33	0.25
	1 / 2	Yes	Predications	413	12	2	22	0.08	0.17
	1	Yes	Predications	129	29	8	87	0.08	0.28
	All Phases			**6444**	**636**	**81**	**195**	**0.29**	**0.13**
Co-occurrence	Marketed	Yes	CoOccSen	35706	919	31	4	0.89	0.03
	3	Yes	CoOccSen	2614	172	30	22	0.58	0.17
	2	Yes	CoOccSen	4723	205	35	35	0.5	0.17
	1 / 2	Yes	CoOccSen	3875	40	7	17	0.29	0.18
	1	Yes	CoOccSen	1342	97	18	77	0.19	0.19
	All Phases			**48260**	**1433**	**121**	**155**	**0.44**	**0.08**

Note: Drug filters were applied across both models, and not other filter was used, so that a direct comparison could be made.

Discussion

SemRep relies on domain knowledge (UMLS) to extract relationships, by applying NLP rules at the sentence level.(Rindflesch & Fiszman, 2003) Normally, the recall is expected to be higher with co-occurrence than SemRep predications, since the former is not restricted by the constraints that the latter imposes on drug-gene pairs. On the other hand, since SemRep is optimized for precision, its results are expected to provide higher precision than co-occurrence. My results are consistent with both of these expectations. As we move from predications to co-occurrence data, a drop in precision, from 0.13 to 0.08 is observed. In contrast, recall increases from 0.29 to 0.44. The increase in recall is more

prominent in lower phase drugs than marketed drugs (0.3 for marketed drugs vs. 0.11 for phase 1 drugs).

In the final section of this chapter I will discuss the implications of these finding, and explain why they indicate a need for further research in this area.

Conclusion and next steps

In this chapter I introduced AIMED system that uses ontology-derived semantic relations as well as co-occurrence statistics to find drugs that interact with genes of interest for the purpose of supporting precision oncology. I found that relying solely on a knowledge-driven system (such as the SemRep NLP system (Rindflesch & Fiszman, 2003)) presented us with two problems. The first problem involved an underrepresentation of oncology drugs in the SemMedDB database (Kilicoglu et al., 2012) due to missing concepts from the underlying ontology. To address this issue, I developed SemMedDB_UTH ("SemMedDB_UTH Database Outline," n.d.), which was constructed by modifying the vocabulary used by SemRep when extracting knowledge from PubMed abstracts. While this improved performance compared to the original edition of the database, I was still faced with the second problem, where SemRep did not recognize some of the relationships of interest, even when the concepts involved were already identified (SemRep relies on the UMLS Semantic Network (At, 1989) to decide which relationships are permissible for any given pair of concepts). To address this issue, which was more prominent in cases where available knowledge was particularly scarce (e.g., drugs in early phase clinical trials), I

used co-occurrence statistics to improve performance. However, recall improved, but precision decreased, since results were no longer constrained by the underlying ontology.

These experiments revealed an underexplored area between the linguistic rules and semantic constraints that systems such as SemRep impose on the one hand (thus achieving higher precision), and the unconstrained relationships defined by co-occurrence (evident by higher recall) on the other. Absence of predefined relationship types to constrain Boolean retrieval can lead to overwhelmingly large result sets. The question arises as to whether other mechanisms than semantic predications (or NLP-based sentence-level relationship extraction in general) might be used to constrain the large numbers of drug/gene co-occurrence instances detectable in the literature to identify drugs of interest. In the following chapter I evaluate the extent to which methods of distributional semantics can be applied to this end.

Chapter 4: Comparing models of attributional and relational similarity for recovery of held-out drug/gene relationships

In the previous chapter, I introduced AIMED, a QA system that tries to find relevant drug-gene relationships for precision oncology, by using knowledge-based NLP methods and unconstrained co-occurrence information. I showed that NLP methods, which depend on established knowledge, have limited coverage in rapidly evolving domains such as precision oncology, and in particular with drugs in lower development phases (evident by low recall). On the other hand, using co-occurrence as a means to find relationships in an unconstrained fashion, presents us with a different problem, as the number of results returned by the system can be too large to be useful (low precision). One potential solution involves statistical systems that neither rely on explicit assertions (co-occurrence), nor are limited to pre-defined relationship types (such as "INHIBITS" in the case of knowledge-based predications), and reason on the basis of *similarity*. In this chapter I explore the utility of a corpus-based approach to this problem, by applying a range of relational and attributional similarity methods, in the framework of the specific aims for my dissertation. Much of the material presented in this chapter is borrowed from a manuscript under review for publication at the time of this writing.

Background

According to the "distributional hypothesis" in linguistics (Harris, 1954) words that occur in similar contexts are likely to have similar meanings. Methods of distributional semantics derive similar representations for terms that occur in similar contexts in the literature.(Trevor Cohen & Widdows, 2009) Thus, two drugs that exist in similar contexts (e.g. a document, or a *sliding window*) may be similar in some respects. *Attributional similarity* concerns the similarity between entities (such as two drugs) (Medin et al., 1990), which with distributional methods is estimated based on the contexts in which they occur across a large corpus. (Landauer & Dumais, 1997; Turney, 1997) In contrast, *relational similarity* concerns the similarity between pairs of entities (such as two drug-gene pairs) (Turney & Littman, 2005), and with distributional methods is estimated from the contexts in which these entity pairs occur together (see for example (Turney, 2005)). While it seems intuitive that relational similarity could help to identify relationships of interest between biomedical concepts, little was understood about the relative merits of relational and attributional similarity as a means to accomplish this task at the outset of this doctoral work. To address this gap in the literature, the work described in this chapter involves an evaluation of the relative utility of relational and attributional similarity for a task of this nature. Specifically, I compare the performance of multiple relational and attributional similarity methods on the task of identifying drugs that may be therapeutically useful in the context of particular molecular aberrations, compared to a gold standard ("the reference set") created by a team of human experts. I use known examples from the reference set as

seeds and apply similarity measures to find *target* drugs in the *search space*. My hypothesis is that relational similarity will be more effective than attributional similarity when applied to this task.

In the sections that follow I will describe the steps that I took to evaluate this hypothesis. I will provide a brief account of the construction of the search space for target drugs, followed by a description of the reference set, and detailed account of the methods used to estimate attributional and relational similarity.

Search space ("Training Corpus")

I used Medline abstracts as the source of information for all similarity models in this evaluation. Specifically, I used additional components of SemMedDB_UTH 2015 (Fathiamini et al., 2016; "SemMedDB_UTH Database Outline," n.d.) (introduced in chapter 3), which provides all the sentences (144M) extracted from 23.4M Medline abstracts dated up to Sep 2014, as well as a list of the UMLS and EntrezGene concepts found in each sentence, their semantic types, and CUIs. I replaced the narrative descriptions of all concepts extracted by MetaMap from the abstracts with their CUIs, and removed stop words using the stopword list from the SMART information retrieval system. (Salton, 1971) For example, "*Sialyl-Tn antigen expression was studied immunohistochemically in 211 primary advanced gastric carcinomas.*" was transformed to "*C0074480 C0185117 studied C1441616 211 C1335475*". I will refer to text so transformed as *CUI-transplanted* text for the remainder of the chapter. The result of this process was a set of 23,610,369 abstracts, with 4,288,491 unique terms, which were

retained in an Apache Lucene index ("Apache Lucene," n.d.) to facilitate search and retrieval. To extract explicit drug-gene pairs and their intervening terms, I further processed individual sentences from the CUI-transplanted abstracts, and whenever a drug co-occurred with a gene in a sentence I extracted the words that lay between them. In this fashion, I identified 52,465,681 drug-gene pair co-occurrence events, and combining their intervening terms (including other CUIs and non-CUI terms) resulted in representations for 6,899,439 unique pairs, each with a "bag of words" (BOW) consisting of every term that occurred between their constituent CUIs in any sentence in the corpus.

Search Space Filters

Methods of distributional semantics produce continuous estimates of relatedness between entities, and as such, they are well suited toward rank-ordering vectors within a search space of potentially therapeutic agents. To construct this search space I removed from the list of extracted concepts any entity that was neither a gene nor a drug. I retained only concepts with UMLS semantic types *aapp, antb, clnd, horm, imft, nnon, opco, orch, phsu* for drugs, and *aapp, enzy, gngm* for genes (*aapp* was used for both drugs and genes), informed by results produced by different configurations of AIMED (Fathiamini et al., 2016). Next, since the goal of the system was to find *clinically relevant* drugs, I used several filters, developed during the course of the AIMED project, to eliminate concepts that met the semantic type constraints, but were not clinically applicable. Specifically, the *NCI drug filter* only includes drugs that are mentioned in the NCI terminology as a "Pharmacologic Substance", the *CT filter* includes drugs mentioned in the clinicaltrials.gov

database, and the *FDA filter* includes only FDA approved drugs. The retrieved entries had to exist in *either* the FDA *or* CT list, *and* the NCI filter to pass the drug filter. To ensure that the performance of pair and entity-based models was compared across the same search space, only drugs and genes that had representatives in both the entity-based *and* pair-based spaces were retained. To meet this last constraint, a drug would need to co-occur at least once with the gene in question. *Figure 4* shows a high-level data flow diagram providing an overview of the data sources and algorithms employed.

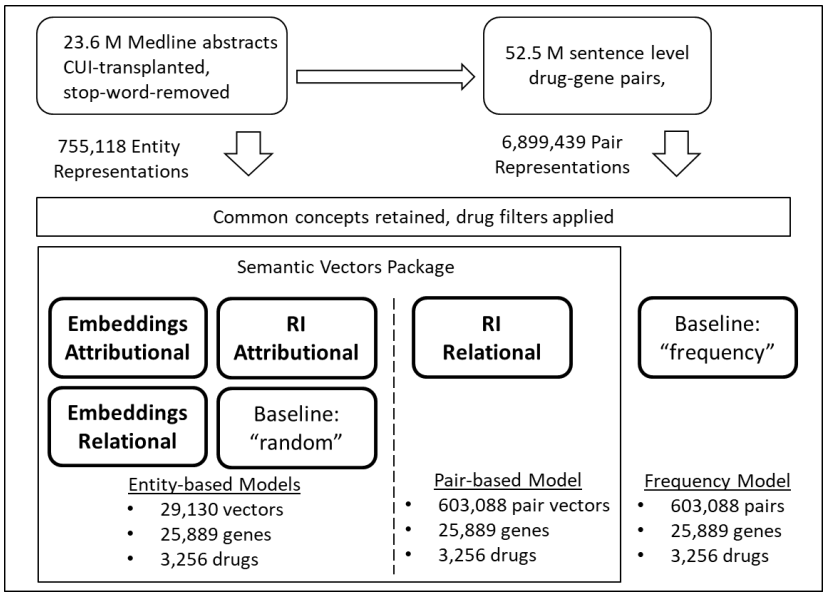

Figure 4. High level data flow diagram from Medline abstracts to different models.

Note: RI=Random Indexing. CUI-transplanted Medline abstracts were used to create entity and pair representations. The drug filters were applied, and only concepts that had representatives in both spaces were retained. The open source Semantic Vectors package (see below) was used to create different vector models: RI Attributional (see below), RI Relational (see below), Embeddings Relational (see below), and Embeddings Attributional (see below). Two other models, "random", and "frequency" (see below) where built to establish a baseline for comparison.

The reference set

As a reference set to test the system output and validate the results, I used the knowledge base provided and maintained by cancer biologists and clinicians at the Precision Oncology Decision Support (PODS) team at the MD Anderson Cancer Center, Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy (IPCT), accessible with permission at http://personalizedcancertherapy.org. Each gene and its associated drugs (collectively known as a Gene Sheet – GS) included in this knowledgebase was deemed by the PODS team to have treatment implications for certain cancer types. To build upon my findings from AIMED, I used the same Gene Sheets from the "evaluation set" of AIMED to test my hypothesis. This list included 17 genes (and some of their synonyms/CUI/Entrez_ID variations), and 430 drugs known to target them (and 1035 synonyms/CUI variations).

All the entries in this reference set were normalized to UMLS CUIs or EntrezGene IDs for genes, henceforth collectively referred to as CUIs in this chapter for uniformity, by SemRep_UTH. Some of the drugs were excluded from the evaluation, either because they were not identified as 'drug' by SemRep_UTH; or because they were not found in the drug filters (explained above). Also, following the practice explained in (Chiu, Crichton, Korhonen, & Pyysalo, 2016), if a drug had no representation in the search space, I systematically disregarded it in the evaluation. This resulted in the GS for one gene (*KIT*) being removed from the reference set, as all its drugs were eliminated in the filtering process. Eventually, 16 genes and 163 drugs were included in the evaluation. Table 1 shows

a list of the genes used for this purpose, and the number of therapeutically-relevant drugs for each of them with representation in my entity-based vector spaces before and after imposition of the constraint that only drugs co-occurring with genes in a CUI-transplanted sentence at least once were included in the evaluation. That is to say, the current experiments, only drugs that met the co-occurrence constraint after filtering (bottom row of Table 9) were considered as positive examples. This co-occurrence constraint is a prerequisite to comparison between pair- and entity-based methods. However, it greatly constrains the number of drugs under consideration, a limitation I will subsequently discuss. This reduction in the number of therapeutically relevant drugs that could be considered for my experiments with the imposition of the co-occurrence constraint had a corresponding effect on the number of drugs remaining in the search space, reducing a total of 3,256 represented drugs (after filtering) to 1,144. The proportion of drugs that were therapeutically relevant in at least one context was similar before (.073) and after (.087) this filtering.

Many drugs that met the constraints for inclusion in the resulting reference set were shared among two or more genes. That is to say, they were considered to be therapeutically active in the presence of an aberration to multiple genes. Out of the 16 genes in this set, five had all their drugs shared with other genes, and only one gene (SMO, targeted by only one drug) shared no drug with the others. *Figure 5* shows a summary of the drug overlap between any given gene and the rest of the genes. An important implication of this overlap is that sets of seed drugs (or seed drug-gene pairs) drawn from other gene sheets may, at

times, include positive examples from the held-out gene sheet used at a particular point in the cross-validation procedure.

Table 9. *List of genes and number of drugs used as the reference set for evaluation*

Gene	ABL1	AKT1	ALK	BRAF	CDK4	CDK6	EGFR	ERBB2	FGFR1	FGFR2	FLT3	KDR	PDGFRA	RET	ROS1	SMO	**Sum**	**Total unique drugs**
Number of therapeutically relevant drugs (TRD)	17	43	5	24	14	7	41	53	29	19	30	53	32	16	3	8	**394**	**237**
TRD found in entity-based spaces (ri_att, emb_att, emb_rel, rand-vec)*	11	23	3	10	4	3	22	26	15	10	19	33	24	11	2	1	**217**	**118**
TRD-gene pairs found in pair-based spaces (ri_rel, frequency)*	9	21	2	10	4	3	20	20	14	10	7	25	12	3	2	1	**163**	**99**

Note: There are fewer representations of drug-gene pairs than there are of therapeutically relevant drugs, as some therapeutically-relevant drugs did not co-occur with the gene in question, prohibiting the generation of a drug-gene pair representation. **Sum**: total number of therapeutically-relevant drug/gene pairs. **Total unique drugs**: total number of drugs that were considered therapeutically relevant in at least one context.
* A detailed description of the models is presented in the following sections

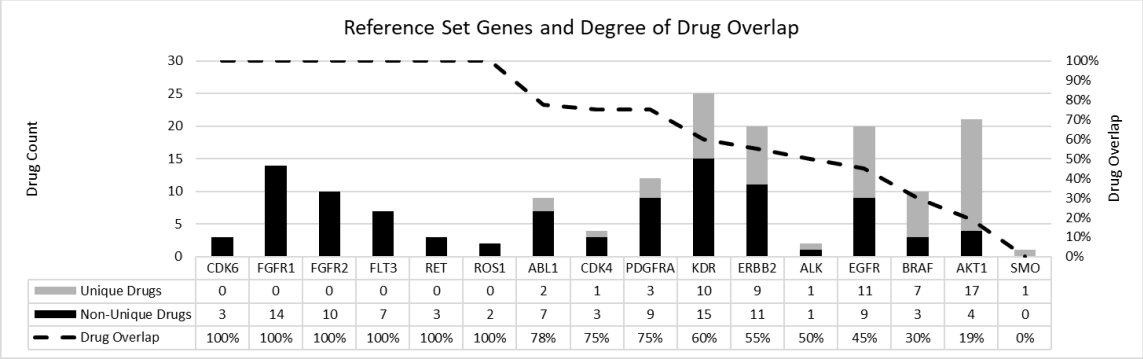

Figure 5. Reference set genes and the percentage of drugs that they share with other genes

Search and Evaluation Process

I used known examples from the reference set as *seeds* and applied similarity measures to find *target* drugs in the search space, and the results were compared against the reference set. Based on the observation that in biomedicine there is often more than one correct answer to any given analogy question (Newman-Griffis, Lai, & Fosler-Lussier, 2017), and since distributional methods aim to prioritize results based on a continuous measure of similarity, I used standard ranked retrieval metrics to evaluate the results. The Average Precision is defined as:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times IsRelevant(k)}{TR}$$

where n = number of results returned

IsRelevant = 1 for therapeutically-relevant drugs, otherwise 0

TR = total number of relevant answers (whether they are returned or not)

P(k) = precision at the point at which the *k*th result was returned.

I also calculated Mean Average Precision (MAP) as the arithmetic mean of the AP values. The details and scope of the models involved in these evaluations are presented in subsequent sections.

Aim 1: Develop and implement models of attributional and relational similarity

In Aim 1, I built models of attributional and relational similarity to test my hypothesis. I used variants of Random Indexing (M. Sahlgren, 2005) and neural word embedding

techniques (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) to build my vector spaces. These operations were performed using the open source Semantic Vectors package[2] (Widdows & Cohen, n.d.; Widdows & Ferraro, 2008) which provides implementations of both of these approaches, eliminating the possibility of introducing bias on account of differences in pre-processing and tokenization of text (*semanticvectors*, n.d.; Widdows & Cohen, n.d.; Widdows & Ferraro, 2008).

Relational similarity models

I used two approaches to model relational information. In the first, I *explicitly* identified drug-gene pairs, and created vector representations for them based on the terms that lie between them when they co-occur in my corpus of CUI-transplanted abstracts. Relational similarity was estimated based on the similarity between these *pair vectors.* A disadvantage of this approach is that all drug-gene pairs must be identified beforehand.

In contrast, in the second approach, I used the *implicit* relational information captured during the course of generating neural word embeddings, and performed geometric operations on the resulting *concept vectors* ($\overrightarrow{Drug_{cue}} - \overrightarrow{Gene_{cue}} + \overrightarrow{Gene_{target}} \cong ?$, as in the example $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Italy} \cong \overrightarrow{Rome}$) (Mikolov, Chen, et al., 2013). Relational similarity was estimated as the cosine metric between the vector resulting from these

[2] https://github.com/semanticvectors/semanticvectors

arithmetic operations and the vector for each drug in the search space (as this will be high if $\overrightarrow{Drug_{cue}} - \overrightarrow{Gene_{cue}} \cong \overrightarrow{Drug_{target}} - \overrightarrow{Gene_{target}}$).

Attributional similarity

To model attributional information, I used CUI-transplanted abstracts as *documents* to build vector spaces, and measured the cosine similarity between *concepts*. Drugs known to be effective against particular genes were used as seeds to find other drugs by assessing their cosine similarity. In my first approach, I used Random Indexing to build the vector space, and the second approach I used the same neural concept embeddings space from the relational similarity experiment, but instead of using relationships, individual drugs were used as seeds to find similar drugs.

Preliminary Experiments and Parameter Selection

Each of the methods introduced above can be executed using different sets of parameters that could affect performance. Preliminary experiments were performed to choose the optimal set of parameters for each model. All models used a minimum word frequency of 10. The vector dimensionality was 1000 for RI-based models (which tend to require relatively high dimensionality), and 500 for neural embedding models (which have been shown to perform well at relatively low dimensionalities).

Attributional similarity with Random Indexing (`ri_att-RI`)

In my first approach, I built a simple Random Indexing (M. Sahlgren, 2005) space. A set of random vectors, one for each document in the corpus was generated by creating zero

vectors of dimensionality 1000 and randomly assigning 10 of these values to either +1 or -1. The result is a set of document vectors with a high probability of being orthogonal, or close-to-orthogonal, on account of the statistical properties of high-dimensional space (M. Sahlgren, 2005). Term vectors were built by adding together the document vectors they occurred in. This process can be expressed as:

$$\vec{T} = \sum_{(t)\in D} randVec(D)$$

where $\vec{T}$ represents the term vector, *D* is a given document, *t* denotes a given term in the document, and *randVec* is the function to assign random vectors to documents.

Attributional similarity with Reflective Random Indexing (`ri_att-RRI`)

In this approach, a Term-based Reflective Random Indexing (TRRI) (Trevor Cohen, Schvaneveldt, & Widdows, 2010) space was built. In TRRI, random vectors are assigned to terms (a combination of terms and CUIs in my case), and added together to generate *document* vectors for documents containing those terms, which are subsequently normalized. Log entropy was used as the term-weighting scheme. This is the beginning of an iterative training procedure – new term vectors are generated by adding together the document vectors for documents in which they occur in, then the cycle can be repeated if necessary. This provides a computationally convenient way of estimating the relatedness between terms that do not co-occur directly together in documents, as terms that co-occur with similar *other* terms will also have similar vectors. The process can be expressed as the following sequence:

1. $\vec{D} = \sum_{(t)\in D} randVec(LE(t))$
2. $\vec{T} = \sum_{(\vec{D}\ \{t|t\in D\})} \vec{D}$

Summarized as:

$$\vec{T} = \sum_{(\vec{D}\ \{t|t\in D\})} (\sum_{(t)\in D} randVec(LE(t)))$$

where $\vec{D}$ represents the document vector, *D* is the set of terms in each document, *t* denotes a given term in the document, *randVec* is the function to assign random vectors to terms, *LE* is the log entropy term weighting function, and $\vec{T}$ is the final term vector. `Ri_att-RRI` was built with the same dimensionality and number of random values as the previously discussed `ri_att-RI` space, over a single iteration (random term vectors → document vectors → term vectors).

Relational similarity with pair vectors and Random Indexing (`ri_rel-RI`)

As a relational counterpart to `ri_att-RI` above, I created vector representations of drug/gene pairs in accordance with the RI paradigm (Kanerva et al., 2000). I treated each distinct BOW (see above) as a *pseudo-document*, generating pair vectors by adding together the random vectors for the terms in each BOW and normalizing the result. No term weighting scheme was used. This process can be expressed as:

$$\vec{P} = \sum_{(t)\in P} randVec(t)$$

where $\vec{P}$ represents the pair (*pseudo-document*) vector, *P* is the set of terms in each BOW, *t* denotes a given term in the BOW, and *randVec* is the function to assign random vectors to terms.

Relational similarity with pair vectors and Reflective Random Indexing (`ri_rel-RRI`)

This model was similar to `ri_rel-RI` in that I treated each distinct BOW as a *pseudo-document*, and created pair vectors by adding together vectors for terms in each BOW and normalizing the result. The difference, however, was that instead of using random vectors for terms, I used the term vectors trained in the process of TRRI for `ri_att-RRI` model explained above. I hypothesized that doing so would provide the means to assess the similarity between pair-based *pseudo-documents* containing semantically related but non-identical terms. The process of generating pair vector representations can be expressed as:

$$\vec{P} = \sum_{(t)\in P} ret\_attrib\_RRI(t)$$

where $\vec{P}$ represents the pair (*document*) vector, *P* is the set of terms in each BOW, *t* denotes a given term in the BOW, and *ret_attrib_RRI* is the function responsible for retrieving term vectors from the `ri_att-RRI` space.

Relational similarity with concept embeddings (`emb_rel`)

A second class of relational models were built using the Semantic Vectors implementation of the Skipgram-with-Negative-Sampling (SGNS) algorithm, following the descriptions

provided in (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) for word embeddings, with the source abstracts (rather than sentences) as documents. With SGNS, a neural network is trained to predict the terms surrounding an observed term, within a sliding window that is moved through the text. The probability of a surrounding term given an observed term is estimated as the sigmoid function of the scalar product between the input weights of the observed term, and the output weights of the surrounding term. The network is trained using stochastic gradient descent to optimize the following objective:

$$\sum_{(t_o,t_c)\in D} \log \sigma\big(I(\mathrm{t_o}).O(\mathrm{t_c})\big) + \sum_{(t_o,t_{\neg c})\in D'} \log \sigma\big(-I(\mathrm{t_o}).O(\mathrm{t_{\neg c}})\big)$$

where D is a set of observed terms (t_o) and their context terms (t_c), D' is a set of observed terms (t_o) and corresponding randomly drawn terms ($t_{\neg c}$) that are unlikely to occur in the context of the observed terms. I denotes the input weights for each term, O denotes the output weights for each term, and σ is the sigmoid function, which converts the scalar product of the input and output weights concerned into a value between 0 and 1 that can be interpreted probabilistically. Optimization of this objective results in high predicted probabilities for terms that occur in the context of an observed term, and low predicted probabilities for terms that do not. The input weights (I) are retained as the word (or concept) embeddings, although it has been shown that retaining the output weights (O) is advantageous in some experimental settings (Levy et al., 2015). Neural embeddings have been shown to capture a form of implicit relational similarity, which can be used to solve proportional analogy problems of the form "***a*** is to ***b*** as ***c*** is to ***what***?" (Mikolov, Yih, et

al., 2013), using simple geometric operations. With this model, assuming *drug1* has a similar effect on *gene1* to *drug2's* effect on *gene2*, an equation can be established such that, assuming relational information is captured accurately: $\overrightarrow{drug1} - \overrightarrow{gene1} \cong \overrightarrow{drug2} - \overrightarrow{gene2}$. In my example, *drug1* and *gene1* are in a known relationship with each other, and the goal is to find *drug2* in relationship with the gene of interest, *gene2*. As such, *drug2* can be found using this equation:

$$\overrightarrow{drug2} \cong \overrightarrow{drug1} - \overrightarrow{gene1} + \overrightarrow{gene2}$$

Attributional similarity with concept embeddings (`emb_att`)

In this experiment, I used the same word embeddings space as the previous model to find drugs similar to known drugs from the reference set.

Parameter variations with embeddings models

Prior work has evaluated the effect of neural word embedding hyper-parameters on task performance in the biomedical domain (Chiu et al., 2016; Levy et al., 2015). I assessed two of those parameters: subsampling (`ss`: the process of ignoring instances of frequently occurring terms with some probability – I used $1 - \sqrt{(T/F)}$ as described in (Mikolov, Sutskever, et al., 2013), where T is a threshold, and F is the number of times a term occurs in the corpus divided by the total number of terms in the corpus) at thresholds of 10^{-3} and 10^{-5}, and window size (`ws`: the number of words considered before and after the target word, in the context of a sliding window) at levels 5 and 8. Furthermore, based on the findings by Levy et al. (Levy et al., 2015) who showed that adding context vectors to word

vectors (`w+c`) with SGNS could help improve performance on pairwise analogy tasks, I tested models with and without context vectors. Overall, six versions of the embeddings search space were built using different combinations of these parameters, as summarized in Table 10.

Baseline models

To establish a baseline and to assess the effect of co-occurrence alone without any similarity measure, the original drug-gene pairs that were identified in the course of building the `ri_rel` models were sorted based on their frequency of co-occurrence across the entire search space. In this model (henceforth: "`frequency`" model), the more a drug co-occurred with a gene, the higher it ranked. For each gene of interest, the resulting ranked list of drugs was compared with the reference set for evaluation.

A second baseline model was built using a set of random vectors for individual concepts (henceforth: "`rand-vec`" model). In a manner similar to the attributional methods described above, drug vectors were used to find similar drugs, and the results were compared with the reference set for evaluation. The intuition here was since the vectors used in this model were randomly chosen, they have a high probability of being orthogonal or close-to-orthogonal to each other. Consequently, any performance observed must occur on account of random overlap between vectors (as they are not perfectly orthogonal), or because drugs overlap across reference sets (as discussed above). Thus, inclusion of the `rand-vec` model permits us to estimate the extent to which observed performance

exceeds that produced by incidental overlap. *Table 10* summarizes different models, and their variants, used for search.

Table 10. *Similarity models used for search.*

	Attributional	**Relational**
Random Indexing	**`ri_att`**: Abstracts as documents, cosine similarity measured between *term vectors* - `ri_att-RI`: term vectors sum of random document vectors (RI) - `ri_att-RRI`: term vectors sum of document vectors trained on random term vectors (TRRI)	**`ri_rel`**: Drug-gene pairs-based BOW as *document*, cosine similarity measured between *pair (document) vectors* - `ri_rel-RI`: document vectors sum of random term vectors (RI) - `ri_rel-RRI`: document vectors sum of term vectors from `ri_att-RRI`
Word Embeddings	**`emb_att:`** Abstracts as *documents*, cosine similarity measured between *term vectors* - `emb_att-001_ws5`: ss=10^{-3}, ws=5 - `emb_att-001_ws8`: ss=10^{-3}, ws=8 - `emb_att-00001_ws8`: ss=10^{-5}, ws=8 All three variations above with `w+c` - `emb_att-001_ws5_w+c` - `emb_att-001_ws8_w+c` - `emb_att-00001_ws8_w+c`	**`emb_rel`**: Abstracts as *documents*, cosine similarity measured after *geometric operations on term vectors*: $\overrightarrow{Cue\ Drug} - \overrightarrow{Cue\ Gene} + \overrightarrow{Target\ Gene} = ?$ - `emb_rel-001_ws5`: ss=10^{-3}, ws=5 - `emb_rel-001_ws8`: ss=10^{-3}, ws=8 - `emb_rel-00001_ws8`: ss=10^{-5}, ws=8 All three variations above with `w+c` - `emb_rel-001_ws5_w+c` - `emb_rel-001_ws8_w+c` - `emb_rel-00001_ws8_w+c`
Baseline	• **`frequency`**: drug-gene pairs sorted by the number of occurrence in the abstracts, search by gene returned drugs • **`rand-vec`**: Abstracts as *documents*, cosine similarity measured between *random term vectors*	

Aim 2: Recovery of held-out drug/gene relationships

In this phase, I evaluated the models from Aim 1 for their ability to recover held-out drugs and drug-gene pairs by using a set of seeds examples from the reference set (introduced previously), across a range of cross-validation configurations.

Cross-validation configurations

As explained previously, both relational and attributional models require seed examples, so that ranked retrieval of target entries can occur based on similarity to these seeds. For attributional models the seed and target were drugs, and for relational models they were drug-gene pairs. With the `frequency` model the "seed" was just the gene in question, and I ranked the drugs that co-occurred with it based on frequency. To evaluate the pair-based models, rankings of retrieved pairs containing the reference set drugs were considered. For the sake of uniformity, I will refer to pair-based seeds and targets, simply as "drugs". I conducted my evaluation both at a single GS level (*InGene* – all cues and targets directly concerned the gene of interest), and across all the GSs (*ExGene* – the gene of interest served as the target, where all the other genes were used as seeds). My hypothesis was that the *InGene* configuration would elicit the best performance from attributional models (as retrieved drugs would be similar to drugs that are known to be effective), while the *ExGene* configuration would elicit best performance from relational models (as the nature of the relationship between therapeutically relevant drugs and the genes they target may be consistent across genes).

InGene models

In the *InGene* model the scope of the cross validation was limited to one single GS at a time (given knowledge of some drugs known to affect *this* gene, can I find others?) I used two cross validation strategies. Both strategies are forms of leave-one-out cross-validation, but they differ with respect to the number of drugs that are retained as seeds. With the first strategy, known as One-As-Seed ("*oas*"), I took one "target" drug at a time from the reference set and used all the other drugs *individually* as seeds to find it and calculate AP. Of note, since there was only one target drug to find, the AP was equivalent to reciprocal rank in this case. MAP for each gene was calculated by averaging the set of AP results (or rather, reciprocal ranks) obtained in this process. For *t* target drugs and *s* seed drugs, the number of reciprocal ranks averaged is *t***s*. The utility of each possible seed for retrieval of each target is evaluated. The second strategy, known as All-But-One ("*abo*"), involved using all the drugs (with vectors combined) to find a single held out drug. In this model the cue was the normalized superposition of the vector representations of all the cues concerned. For each gene, MAP was then calculated across this set of AP results (or more accurately, reciprocal ranks) (one for each held-out drug). Irrespective of the number of seed drugs, this average was calculated over *t* reciprocal rank results. As such, the main difference between "*oas*" and "*abo*" was that in the former, seed drugs were used *individually* to find the target drug with the results averaged later, whereas in the latter, a *cumulative* seed vector was used as a cue. The motivation for this design was that in emerging domains, a single positive result could be useful as a means to identify other

results (as in the case where annotators have yet to begin constructing a gene sheet), and building the basis for further discoveries – hence the *oas* model. On the other hand, when information is already available (as in the case of an existing gene sheet that needs to be maintained as new potentially useful drugs are described in the literature), one would try to maximize the robustness of the query vector by including in it as many existing positive answers as possible – hence the *abo* model. It has been shown that combining multiple examples as cues lead to better performance on analogical reasoning experiments. (Trevor Cohen et al., 2011; Drozd, Gladkova, & Matsuoka, 2016) As such, my hypothesis was that in any given class of experiments, the *abo* models would perform better than *oas*.

ExGene models

In the case of *ExGene* model (given knowledge of drugs known to affect *other* genes, can I find those affecting this one?), the *oas* model was implemented by first adding (and normalizing) the vectors for individual drugs under each seed gene to form one *prototypical* drug vector for each GS (one *gene sheet* as seed), and then using that vector to find the drugs that target the target gene. Consequently, with *t* target drugs for a gene sheet, and *g* other gene sheets, the MAP was calculated by averaging across *g* average precision results. With *ExGene*, the *abo* model simply involved adding up the vectors for all the drugs under all the seed genes (and normalizing them afterwards) to use as the seed. Consequently, with *t* target drugs for a gene sheet, the MAP simply equaled the average precision, which was calculated only once per target gene, irrespective of the number of other gene sheets or *t*. Figure 6 shows a diagram of the cross-validation configurations. I tested the models

described in *Table 10* with these cross-validation configurations, and report the median of MAP values for the genes in the reference set. Also, as explained previously, many genes in the reference set had drugs that were also mentioned in other Gene Sheets. I hypothesized that this drug overlap would affect the MAP results for ExGene models, since for those genes, seed and target sets have drugs in common. Positive correlation between model performance and the degree of drug overlap may explain the results. To this end, I ran a Spearman Rank Order test to evaluate the correlation between degree of drug overlap among genes in the reference set, and the MAP results for each gene-model combination.

Final Filtering of Result

In all of the evaluations explained above, a drug-gene co-occurrence filter was applied to each result set from entity-based models, before calculating the AP. For each such model, drugs that did not co-occur with the gene in question in at least one original source sentence were eliminated, so that entity and pair-based models could be compared against the same set of constraints.

Results

I ran some preliminary experiments to determine the best set of hyperparameters for the models. A summary of the net effect of those hyperparameters on model performance is presented in Table 11. Based on these findings, I chose the following model configurations as the representatives in their respective categories: `ri_att-RRI` for attributional RI, `emb_att-001_ws5_w+c` for attributional embeddings, `ri_rel-RI` for relational RI, and

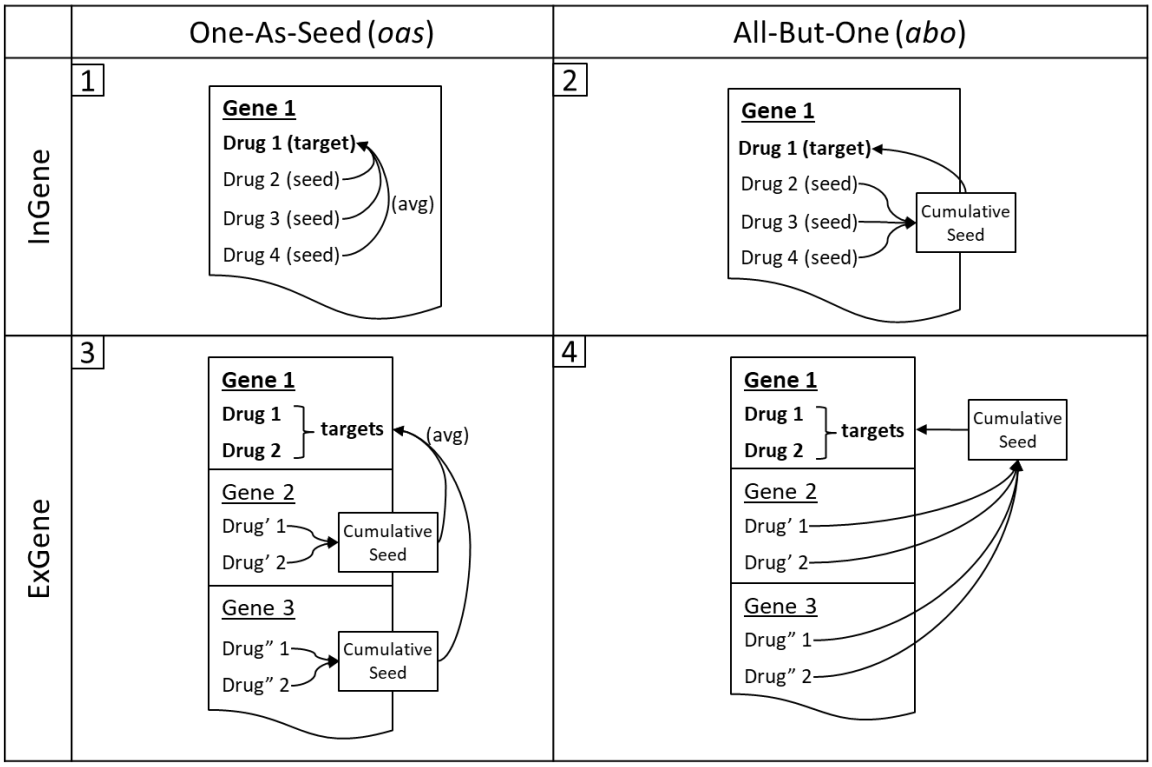

Figure 6. Diagram of different cross validation models.

Note: 1) *oas-InGene*: Drugs in a Gene Sheet are used individually to find a target drug, 2) *abo-InGene*: Drugs in a Gene Sheet are combined (vectors superimposed, normalized), and used to find a target drug, 3) *oas-ExGene*: Gene Sheets are used individually (with drugs within each combined), to find drugs in a target Gene Sheet, 4) *abo-ExGene*: Gene Sheets are used in combination (all their drug vectors combined) to find drugs in a target Gene Sheet. In *oas* models, results from individual queries are averaged (shown as "(avg)" on the diagram) and reported as AP for the target drug(s).

`emb_rel-00001_ws8_w+c` for relational embeddings.

As shown in Table 12 below, the best performing model overall was `emb_rel` abo-ExGene, followed by `emb_rel` oas-ExGene, `ri_rel` abo-ExGene, and `ri_att`

Table 11. *Effect of different hyperparameters on model performance.*

Hyperparameter	emb_rel	emb_att	ri_rel	ri_att
Adding context to word vectors	**increase - 40%**	**increase - 25%**	n/a	n/a
Subsampling threshold from 0.001 to 0.00001	increase - 21%	*decrease - 3%*	n/a	n/a
Window size from 5 to 8	increase - 17%	*decrease - 2%*	n/a	n/a
Replacing RI with RRI	n/a	n/a	*decrease - 23%*	increase - 250%

Note: Average increase/decrease is shown for each model across different configurations (abo/oas, InGene/ExGene). Adding context to word vectors consistently improved performance across embedding models, a finding shown in **boldface**. Some of the hyperparameters resulted in a decrease in performance, shown in *italics*.

abo_InGene. Across RI-based models, ri_rel outperformed ri_att in the ExGene configurations, but not in the InGene categories. With embeddings-based models, emb_rel performed better than emb_att in ExGene models (emb_rel is not defined with InGene). Finally, the *abo* configurations were associated with better performance than *oas* in all models, with only one exception, ri_att oas-ExGene. The results of the correlation test that I performed to assess a potential link between some of the results, and the degree of drug overlap in the Gene Sheets are presented in Table 13.

A brief review of the practical utility of the methods

As discussed previously, a substantial proportion of therapeutically relevant drugs were eliminated to facilitate comparison with pair-based models. To better estimate the practical utility of these approaches I tested the best performing models from the relational and

Table 12. *MAP per gene-model combination, and the median MAP per gene.*

			Median MAP	ABL1	AKT1	ALK	BRAF	CDK4	CDK6	EGFR	ERBB2	FGFR1	FGFR2	FLT3	KDR	PDGFRA	RET	ROS1	SMO
Relational	ri_rel	oas-InGene	0.10	0.12	0.04	0.38	0.12	0.01	0.03	0.09	0.05	0.10	0.11	0.15	0.05	0.17	0.14	**1.00**	0.00
		abo-InGene	0.30	0.38	0.06	0.75	0.31	0.01	0.05	0.33	0.25	0.31	0.21	0.30	0.35	0.75	0.28	**1.00**	0.00
		oas-ExGene	0.34	0.37	0.12	0.63	0.40	0.18	0.29	0.34	0.33	0.36	0.28	0.32	0.29	0.38	0.33	0.66	0.74
		abo-ExGene	0.53	0.54	0.22	**1.00**	0.55	0.36	0.48	0.53	0.48	0.51	0.50	0.53	0.43	0.54	0.56	**1.00**	**1.00**
	emb_rel	oas-ExGene	0.72	0.71	0.38	**1.00**	0.83	**0.64**	0.74	0.65	**0.67**	0.42	0.53	0.34	0.93	0.92	0.73	0.93	**1.00**
		abo-ExGene	**0.75**	0.74	0.39	**1.00**	**0.85**	0.63	**0.76**	**0.75**	0.65	0.35	0.53	0.29	**0.95**	0.88	0.81	**1.00**	**1.00**
Attributional	ri_att	oas-InGene	0.16	0.20	0.09	0.53	0.23	0.27	0.38	0.11	0.09	0.09	0.14	0.14	0.08	0.20	0.52	0.17	0.00
		abo-InGene	0.46	0.69	0.31	0.75	0.68	0.47	0.55	0.47	0.32	0.17	0.27	0.45	0.25	0.58	0.78	0.17	0.00
		oas-ExGene	0.16	0.20	0.09	0.10	0.12	0.09	0.10	0.12	0.15	0.28	0.31	0.32	0.36	0.49	0.55	0.17	0.03
		abo-ExGene	0.14	0.20	0.09	0.05	0.14	0.06	0.05	0.11	0.13	0.40	0.45	0.57	0.47	0.66	0.79	0.15	0.03
	emb_att	oas-InGene	0.18	0.20	0.07	0.49	0.20	0.20	0.21	0.09	0.08	0.11	0.16	0.15	0.14	0.25	0.50	0.67	0.00
		abo-InGene	0.23	0.47	0.15	0.38	0.60	0.21	0.24	0.20	0.12	0.18	0.22	0.20	0.55	0.83	0.56	0.67	0.00
		oas-ExGene	0.40	0.32	0.12	0.28	0.51	0.40	0.32	0.24	0.28	0.39	0.42	0.36	0.87	0.94	0.58	0.75	0.58
		abo-ExGene	0.41	0.30	0.12	0.23	0.66	0.44	0.43	0.25	0.29	0.38	0.39	0.37	0.90	**0.96**	0.67	0.75	0.50
Baseline	frequency		0.35	0.46	0.11	0.70	0.36	0.30	0.52	0.34	0.31	0.15	0.17	0.09	0.32	0.48	0.53	0.83	**1.00**
	rand-vec	oas-InGene	0.02	0.02	0.01	0.03	0.04	0.01	0.00	0.01	0.02	0.02	0.04	0.03	0.03	0.12	0.14	0.00	0.00
		abo-InGene	0.02	0.02	0.01	0.02	0.02	0.01	0.01	0.02	0.01	0.02	0.04	0.01	0.07	0.23	0.28	0.00	0.00
		oas-ExGene	0.15	0.18	0.03	0.12	0.12	0.07	0.08	0.06	0.08	0.23	0.26	0.24	0.27	0.39	0.49	0.19	0.03
		abo-ExGene	0.27	0.45	0.03	0.21	0.23	0.18	0.14	0.06	0.09	**0.69**	**0.59**	**0.73**	0.60	0.73	**1.00**	0.31	0.03
*Included reference drugs				9	21	2	10	4	3	20	20	14	10	7	25	12	3	2	1
†Drugs with vector representations				3,256															
‡Drugs co-occurring with genes				213	797	50	231	160	183	501	302	295	141	240	177	54	24	33	76

Note: Best results for each attributional or relational method are underlined, and best result for each gene sheet and overall are shown in **boldface**.
*Number of drugs in the reference set copied from Table 9.
†Drugs in the vector space after applying filters explained earlier in the text.
‡ Number of drugs available for search per gene concerned. The co-occurrence constraint explained earlier effectively reduced the number of drugs available for search from 3,256 to 1,144 *unique* drugs, with an average of 217 available for consideration for each gene (searchable drugs are shared among the genes).

attributional categories with the full set of available drugs in the reference set (394 therapeutic applications for drugs across 16 Gene Sheets, Table 9) with all the other constraints the same as the main experiment, and found the median MAP to drop an average of 0.26 across those representative models (Table 14). In doing so, I am penalizing the models for not finding drugs that are not represented in the vector space, placing a hard ceiling on performance. It is notable that in this case, the relational models still outperformed the attributional models, a finding consistent with those of the main experiment.

Table 13. *Spearman Rank-Order Correlation Coefficient values*

Model	emb_rel		ri_rel		emb_att		ri_att			rand-vec	
Config	oas-ExGene	abo-ExGene	oas-ExGene	abo-ExGene	oas-ExGene	abo-ExGene	**oas-ExGene**	**abo-ExGene**	frequency	**oas-ExGene**	**abo-ExGene**
Overlap/MAP Correlation	-0.4	-0.39	-0.32	-0.25	-0.03	0	**0.55**	**0.51**	-0.32	**0.6**	**0.63**
MAP	0.72	0.75	0.34	0.53	0.40	0.41	0.16	0.14	0.35	0.15	0.27

Note: The table shows a possible link between genes with high drug overlap, and the MAP values for ExGene configurations. The results are summarized per model. Some of the models show high correlation between their results and the degree of overlap (e.g. rand-vec oas-ExGene and ri_att oas-ExGene) which may help explain their higher-than-anticipated MAP. Further details are discussed in the Discussion section. High correlation values are shown in **boldface**.

Table 14. *Original vs. full reference set.*

Category	Model / configuration	Median MAP Original	Median MAP Full Ref	Drop
Relational	`ri_rel` abo-ExGene	0.53	0.25	0.28
Relational	**`emb_rel` abo-ExGene**	0.75	**0.34**	0.41
Attributional	`ri_att` abo-InGene	0.46	0.15	0.31
Attributional	`emb_att` abo-ExGene	0.41	0.16	0.25
Baseline	`frequency`	0.35	0.16	0.19
Baseline	`rand-vec` abo-ExGene	0.27	0.15	0.12

Note: Effect of moving from using reference drugs that had representatives in the search spaces (Original) to the full reference set irrespective of whether the target drugs were represented in a space or not (Full Ref). Best results for attributional or relational categories are underlined, and best result overall is shown in **boldface**. On average the median MAP drops by 0.26. Only the results for best performing models in each category are shown.

Discussion

My main hypothesis was that relational similarity would be more effective than attributional similarity in finding drugs that interact with particular genes. To this end, for each category of relational similarity, I also developed an attributional counterpart. The results indicate that models based on relational similarity generally outperform models based on attributional similarity on this task, providing strong support for the utility of analogical reasoning (exemplified by relational similarity) in the task of identifying clinically relevant relationships in natural language text.

A related hypothesis was that ExGene configurations would be advantageous for relational models, whereas attributional models may perform best with InGene. This hypothesis was

supported in part by the results, as the Random Indexing based relational model exhibited its best performance in ExGene settings, leveraging relationships involving other genes (I did not compare relational embedding techniques for InGene configurations, as the `emb_rel` model is only defined for ExGene). However, I also anticipated that attributional models would perform worse in ExGene settings (where cue drugs interact with other genes than the target gene). This was exemplified by the `ri_att` model, with a performance drop from a MAP of 0.46 in abo-InGene to 0.14 in abo-ExGene. However, `emb_att` surprisingly displayed the opposite behavior, where its performance improved upon moving from InGene to ExGene (0.23 to 0.41). This paradoxical behavior may be due to the fact that in many cases the genes may be functionally related to one another, a hypothesis that is further supported by the drug overlap among Gene Sheets explained previously. Further investigation is needed to fully explain this phenomenon, as it is not clear why this would occur with one attributional model, but not the other.

A third hypothesis was that *abo* models would generally perform better than their *oas* counterparts. This hypothesis held true across the majority of the experiments (with one exception, `ri_att` oas-ExGene), suggesting that in emerging domains, where existing knowledge is limited, the best strategy for creating robust query vectors may be to use as many existing positive cues as possible. This finding is consistent with previous work on analogical reasoning using distributed representations of semantic predications ("concept relation concept" triples) extracted from the biomedical literature using SemRep (Trevor Cohen et al., 2011), as well as by subsequent work on analogical retrieval in the general

domain (Drozd et al., 2016). As more positive examples are found, their addition to an existing query vector will progressively add to the robustness of the query.

Regarding the nature of the underlying representation, the `emb_rel` model consistently outperformed `ri_rel` both in *oas* and *abo* configurations. The `emb_att` model, however, was only marginally better than `ri_att` with oas-InGene, and in the case of abo-InGene, it fell short of this simpler model. This apparent disadvantage might be due to the context size for the two models. While the `ri_att` model used the whole Medline abstract, `emb_att` only used a small sliding window, which provides a limited scope, and may help explain the poor performance. Further research is needed to test this hypothesis, perhaps by providing a larger window for the neural embedding model, or adapting it to treat entire documents as contextual units.

Another advantage of the `emb_rel` model over `ri_rel` was ease of generation, efficiency, and scalability. Embedding models represent individual concepts as vectors. To create the `ri_rel` search space, I had to first find and extract explicit drug-gene pairs from individual sentences, and then create bags-of-words from their intervening terms, a computationally demanding pre-processing step that took considerable effort to develop, and must be repeated whenever new information is added to the corpus. Furthermore, the resulting vector space is larger as each pair, rather than each entity, must be represented with a unique vector. Given both the level of development, execution effort, and overall performance, the concept-level `emb_rel` model offers clear advantages for relational retrieval.

A surprising finding amongst the results was the performance of the random vector based baseline model (`rand-vec`). I expected negligible performance, as random vectors are by design generated with a high probability of being mutually orthogonal or close-to-orthogonal, and as such are not meaningfully similar to one another. While I obtained the expected results with InGene models, those for ExGene were surprisingly productive, particularly the median MAP of 0.27 for abo-ExGene. I believe this phenomenon is explained by the overlap between drugs across gene sheets, providing the model with same vector both as a seed and as target. This theory is supported by the fact that using the `rand-vec` model, I obtained better results with genes that shared many drugs with other genes than those which did not (e.g., FGFR1, FGFR2, FLT3, KDR, PDGFRA, RET). As shown in Table 13, there is a high correlation between drug overlap and `rand-vec` results in the ExGene category, 0.6 and 0.63 for oas-ExGene and abo-ExGene, respectively. The other baseline model was `frequency`, which I compared to the relational models. While with a median MAP of 0.35, the `frequency` model seems relatively strong in terms of its ability to find gene-related drugs, it outperforms neither `ri_rel`, nor `emb_rel`, indicating that these models are more effective than a simple count of co-occurrence in finding the desired relationships.

Comparison with existing work

The results are not directly comparable to prior work in different domains. The literature is relatively sparse on the application of neural concept embeddings in precision oncology, or even biomedicine, as compared with the general domain. In particular, I am aware of

only one paper in the biomedical domain that concerns using neural word embeddings derived from unstructured text (as opposed to neural embeddings derived from semantic predications (Trevor Cohen & Widdows, 2017)) for analogical retrieval (Newman-Griffis et al., 2017), and this work does not compare attributional and relational models. As mentioned previously, EBC provides an alternative method to `ri_rel` for estimating relational similarity, however it is not directly comparable to my work, since my corpus has not been parsed for grammatical dependencies. Future work, however, includes parsing the corpus to find those dependency paths (or leveraging the set provided by the creators of EBC (Percha, Altman, & Wren, 2018b)) so that EBC can be used. As an attributional counterpart to EBC, Levy and Goldberg's dependency based embeddings (Levy & Goldberg, 2014) can be considered.

Another factor that complicates direct comparison with existing work involves exploration of the space of model hyperparameters, which often resulted in improved performance. Levy et al. provide an extensive description of the set of SGNS hyper-parameters that can be altered to improve the embedding results (Levy et al., 2015). Among the many parameters they explain, I chose to examine three – window size, sub-sampling threshold, and adding context vectors to word vectors. In line with previous work, I found that adding context vectors to word vectors consistently improved word embedding results (across all the cross validation configurations) (Levy et al., 2015). Future work involves performing a more comprehensive experiment to determine the effect of these and other parameters.

Limitations

I faced two problems when dealing with drug-gene relationships in precision oncology. The first problem concerned term to concept mapping (performed by MetaMap), and the other had to do with finding relationships of interest. In the current project, I specifically focused on the latter to fulfil the primary goal of this research – comparative evaluation of different similarity models. Some drugs (119 out of 237 or 50%, Table 1) in the reference set were excluded from evaluation, either because they had no representative in vector space (e.g. because they were not mapped to CUIs by MetaMap,), or because they did not pass the drug filters that I used (which were also based on CUIs). An additional 19 drugs were excluded because of the co-occurrence filter. As such, some true positive results that would have been missed were excluded to allow a "fair" comparison of models.

However, to estimate practical utility, the full reference set should be used. As shown in Table 14, penalizing the models for missing drugs that they do not represent results in a substantial drop in performance. More work is needed to address the limited coverage of therapeutically relevant agents, an issue I hope to address by replacing the concept extraction component of the system in the future. This may involve further expansion of MetaMap vocabularies, or substitution of an alternative method for the recognition of drug and gene entities that is not dependent on curated knowledge resources, which would be advantageous in emerging domains such as precision oncology.

In addition, both the literature and the reference set used in this research were around 2-3 years old. Emerging domains by definition evolve at a rapid pace, and so should the search

spaces and reference sets used in information retrieval research projects in these domains if the resulting systems are to be practically useful.

Furthermore, while I tried to follow the current literature in selecting model hyper-parameters, the current work should not be considered an exhaustive test of these parameters. It is quite possible that other adjustments could further improve performance.

Up to this point, I tested my assumptions and techniques using cross validation across a set constructed by a single team of PODS curators. So, the methods have not been tested in other contexts or for similar tasks. However, the PODS curators constructed the reference standard independently of the computational work and the main goal of this research was to compare different similarity methods and paradigms. In the next chapter I use seed drugs produced by NLP, to test the methods when used with an independent set of cues. This is an important step in terms of evaluating the utility of the developed methods when applied in a more practical scenario.

Chapter 5: Unsupervised identification of clinically relevant drug/gene relationships

In the previous chapter, I compared the utility of a broad range of relational and attributional models for the task of finding relevant drug-gene relationships. In those experiments, both the "seeds" and the held out "answers" came from the same expert-curated reference set. While this type of cross validation can serve to demonstrate the utility of the developed models, it can only be used in cases where some positive cues are already known to the system. In practice, this may not always be feasible. In this chapter I use cues that are extracted automatically from biomedical literature, using NLP, and evaluate the performance of attributional and relational models developed in Aim 1.

Methods

I used SemMedDB_UTH and isolated predications that were associated with the target genes in the reference set. I chose predications with predicate types 'INHIBITS', 'INTERACTS_WITH', 'COEXISTS_WITH' that had any of the target genes as "subject", and any *drug* (concepts with UMLS semantic types *aapp, antb, clnd, horm, imft, nnon, opco, orch, phsu)* as "object". Table 15 shows the number of predications found for each target gene in this manner.

Table 15. *Predications found for each target gene.*

Gene	ABL1	AKT1	ALK	BRAF	CDK4	CDK6	EGFR	ERBB2	FGFR1	FGFR2	FLT3	KDR	KIT	PDGFRA	RET	ROS1	SMO
Pred. Count	299	1,992	116	329	319	286	1,124	778	328	299	289	256	184	87	89	21	74

It must be noted that the predications are not unique to Gene Sheets, as different genes may share certain characteristics, and the same drug may target more than one gene. This is similar to the reference drug overlap phenomenon discussed in the previous chapter, where the unexpectedly high performance of some models (such as `rand-vec`) seemed to be associated with the overlap. Similarly, assertions that are repeated in the context of more than one gene may have a better chance of being accurately extracted.

Next, I tested the best performing models from the relational and attributional categories from Aim 2 with these predications as seeds, and with all the other constraints the same as the main experiment.

Configurations

Unlike the models explained in Aim 2 where *seeds* and *targets* came from the same reference set, in this experiment the seeds (predications from SemMedDB_UTH) were from a different set than the reference set. Therefore, cross-validation (*abo* or *oas*) was neither defined, nor required, as the predications were used as seeds (known as Predications-As-Seed, "*pas*") to find the targets. Arguably though, the *pas* model is more similar to *abo* than *oas*, since in *pas*, one cumulative seed vector, the sum of vectors for all

the predications concerned, is used as cue to find one drug in a single Gene Sheet at a time (*InGene*), or all the drugs in one Gene Sheet (*ExGene*).

Results

The results are summarized in *Table 16*. The best performing model in this experiment was embeddings relational model (MAP of 0.64), followed by the RI-based relational model (MAP of 0.31). Both the relational models outperformed their attributional counterparts. An important question concerned how these *pas* models would compare against their *abo*

Table 16. *Predications used as seeds*

Predications As Seed (PAS)		**Median MAP**	ABL1	AKT1	ALK	BRAF	CDK4	CDK6	EGFR	ERBB2	FGFR1	FGFR2	FLT3	KDR	PDGFRA	RET	ROS1	SMO
Relational	ri_rel pas-ExGene	0.31	0.39	0.12	0.75	0.18	0.09	0.17	0.16	0.32	0.29	0.25	0.30	0.33	0.51	0.63	0.61	0.33
	emb_rel pas-ExGene	**0.64**	**0.53**	**0.32**	**1.00**	**0.68**	**0.57**	**0.59**	**0.70**	**0.60**	0.29	**0.49**	**0.32**	0.87	0.92	**0.81**	**1.00**	**1.00**
Attributional	ri_att pas-ExGene	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.06	0.04	0.05	0.03	0.10	0.15	0.16	0.05	0.01
	emb_att pas-ExGene	0.35	0.24	0.11	0.13	0.39	0.33	0.19	0.21	0.28	**0.37**	0.43	0.25	**0.92**	**0.94**	0.67	0.64	0.50
Baseline	rand-vec pas-ExGene	0.12	0.08	0.04	0.08	0.04	0.28	0.36	0.05	0.18	0.07	0.10	0.14	0.22	0.38	0.27	0.13	0.02

Note: With predications used as seeds, MAP per gene-model combination, and the median MAP per model across all the genes are shown. Best results for attributional or relational categories are underlined, and best result for each gene sheet and overall are shown in **boldface**. Other system parameters including the total number of drugs in the vector space, and the number of drugs in the reference set are identical to the main experiment from Aim 2. This experiment was only run for the best performing models in each category from Aim 2.

counterparts from the cross-validation experiment in Aim 2. The answer to this question could help elucidate the role of NLP (versus human experts) as the provider of cues in this task (although curation of the results would still require human input). As summarized in Table 17, the median MAP dropped an average of 0.13 across the five models when moving from *abo* to *pas*.

Table 17. *Comparing pas models with their abo counterparts*

Category	**Model / configuration**	**Median MAP *abo***	**Median MAP *pas***	**Drop**
Relational	ri_rel ExGene	0.53	0.31	0.22
Relational	**emb_rel ExGene**	**0.75**	**0.64**	0.11
Attributional	ri_att ExGene	0.14	0.03	0.11
Attributional	emb_att ExGene	0.41	0.35	0.06
Baseline	rand-vec ExGene	0.27	0.12	0.15

Practical utility of the methods with predications as seeds

I tested the models with the full set of available drugs in the reference set (394 therapeutic applications for drugs across 16 Gene Sheets, Table 9) with all the other constraints the same as the main experiment. In this case, on average the median MAP dropped by 0.17

across the five models (Table 18). The relational models still outperformed the attributional models, a finding consistent with those of the main experiment.

Table 18. *Full reference set (Full Ref) versus the original configuration.*

Category	Model / configuration	Median MAP Original	Median MAP Full Ref	Drop
Relational	ri_rel pas-ExGene	0.31	0.12	0.19
Relational	**emb_rel pas-ExGene**	**0.64**	**0.27**	0.37
Attributional	ri_att pas-ExGene	0.03	0.01	0.02
Attributional	emb_att pas-ExGene	0.35	0.16	0.19
Baseline	rand-vec pas-ExGene	0.12	0.05	0.07

Note: Best results for attributional or relational categories are underlined, and best result overall is shown in **boldface**.

Summary of the findings

Table 19 summarizes the findings across all the experiments in this chapter. As discussed, the best performing model overall was Relational Embeddings (`emb_rel`) across the four categories of experiments.

Table 19. *Summary of the overall findings.*

	Controlled Reference Set - 16 Genes - 163 Drugs - shared by search spaces - limited to co-occurrence	**Full Reference Set** - 16 Genes - 394 Drugs - all drugs
Supervised: Within Reference Set		
Best model	**Relational embeddings: 0.75** (abo-ExGene)	**Relational embeddings: 0.34** (abo-ExGene)
Best attributional	`ri_att`: 0.46 (abo-InGene)	`emb_att`: 0.16 (abo_ExGene)
Random baseline	0.27	0.15
Unsupervised: Predications as Seed		
Best model	**Relational embeddings: 0.64** (pas-ExGene)	**Relational embeddings**: **0.27** (pas-ExGene)
Best attributional	`emb_att`: 0.35 (pas-ExGene)	`emb_att`: 0.16 (pas-ExGene)
Random baseline	0.12	0.05

Note: Relational embeddings model outperformed both the RI-based relational models (not shown), and the attributional models. The median MAP values are presented for each model.

Discussion

In this chapter, I showed that using the developed methods in an unsupervised manner still produces results that are consistent with my main hypothesis. On the other hand, when testing with the full reference set, the best performing model in this approach had a clear advantage over the randomly created baseline (MAP of 0.27 vs. 0.05 for the random model). This can be important from a practical perspective, since it is an indication of the

relative utility of the methods for this task. The practical utility becomes more important when dealing with a new domain where there is little prior knowledge available. Using the system in those scenarios will provide the curators with some initial cues that will then help build upon them to expand their search, and find more answers in an iterative fashion.

Chapter 6: Contributions, conclusion, and future direction

In the work described in this dissertation, I compared relational and attributional similarity measures for their utility in finding clinically relevant drug/gene relationships in the context of precision oncology, which presents unique challenges on account of the pace of evolution of clinically actionable knowledge. I found that models based on relational similarity outperformed models based on attributional similarity on this task. This finding consistently held true in multiple experiments across the two large paradigms of distributional semantic methods, Random Indexing (RI) (M. Sahlgren, 2005), and neural word embeddings (NWE) (Mikolov, Chen, et al., 2013). This is the first time methods of relational and attributional similarity have been systematically compared in this manner, and as the methods can be applied to identify any sort of relationship for which cue pairs exist, my results suggest that relational similarity may be a fruitful approach to apply to other biomedical problems. Furthermore, I found models based on NWE to be particularly useful for this task, given their higher performance than RI-based models, and significantly less computational effort needed to create them.

In my preliminary work, I developed the AIMED system (Fathiamini et al., 2016) to find relevant drug-gene relationships for precision oncology, using NLP and sentence based co-occurrence. AIMED showed promising results, but it also revealed some of the

shortcomings of knowledge based NLP methods and co-occurrence statistics, especially with early stage drugs, which provided the practical motivation for this dissertation. The current research takes an important step toward a better AIMED application, by providing it with more robust alternative techniques that can potentially address some of its shortcomings.

In the section that follows I will proceed to reevaluate my main hypothesis in the light of the research findings I have documented in the preceding chapters.

Assessment of hypotheses

My main hypothesis was that measures of relational similarity would be of greater utility than attributional similarity for the task of identifying biological relationships that may answer clinical questions in the context of rapidly changing domains. My results provide strong support in favor of this hypothesis, with estimates of relational similarity yielding better performance than comparable measures of attributional similarity across multiple experiments.

Additionally, during the course of these experiments, I developed other hypotheses that were closely related to the main hypothesis. I found out that the best strategy to maximize the robustness of a similarity-based query across a large vector space was to add vector representations of as many cues as possible to construct a query vector. This finding was supported by the observation that my *abo* models (in which all the existing cues would form one cumulative vector to find one held out answer) outperformed my *oas* models

(where cues consisted of only one cue). This finding is consistent with prior research both in the biomedical (Trevor Cohen et al., 2011), and the general domain (Drozd et al., 2016).

A related hypothesis with potential practical implications for search in domains with emerging knowledge is that when looking for drugs that target a gene, using information about the relationships involving *other genes* as cues helps improve the accuracy of system responses in relational models. In fact, relational models actually performed better with cues concerning other genes, than with cues derived from held-out components of the gene sheet under evaluation. This hypothesis was supported by the results showing that the *ExGene* configurations consistently outperformed the *InGene* settings, when used with relational models. From a practical point of view, this finding means that prior knowledge of drug-gene relationships *in general* can facilitate the search of drugs targeting a gene of interest. Thus, relational models have more information to draw upon than attributional models, and the search for drugs targeting a specific gene can proceed without the need for an agent that is already known to be effective to serve as an exemplar.

Theoretical Contribution

Similarity is a fundamental cognitive construct. (Medin et al., 1990) Similar concepts are thought to belong to the same conceptual category in the human mind (Medin et al., 1993), new concepts are thought to be assigned to existing categories based on how similar they are to concepts exemplifying these categories, and evidence suggests that memory relies on similarity operations to retrieve concepts. (Medin et al., 1993) In my experiments, I evaluated methods that leverage mechanisms of analogical retrieval to elicit relational

similarity. This is consistent with cognitive theories of analogy, which suggest that relational similarity is the most important aspect of similarity for analogy processing and retrieval.(HOLYOAK & THAGARD, 1989; Medin et al., 1990, 1993) My results indicate that models based on relational similarity generally outperform models based on attributional similarity in the task of identifying clinically relevant relationships in natural language text, providing strong support for the utility of analogical reasoning for this task. In other words, my work shows that the same mechanisms that have been proposed to explain experimental data on analogical retrieval can also be leveraged for practical tasks in the biomedical domain.

Informatics Contribution

This research compares methods of relational and attributional similarity, using methods of distributional semantics (Trevor Cohen & Widdows, 2009) when applied to finding desired relationships in emerging biomedical domains, and specifically, precision oncology. I used techniques of Random Indexing (M. Sahlgren, 2005) and neural word embeddings (NWE) (Mikolov, Chen, et al., 2013), and was able to establish the latter as the technique of choice for this task across multiple experiments. To the best of my knowledge, the relative utility of relational and attributional similarity for tasks of this nature has not been systematically evaluated in biomedicine previously. Moreover, the utility of NWE-based relational similarity in finding concept pairs using exemplar cue pairs has not been explored in the context of emerging biomedical knowledge in general, and precision oncology in particular.

Practical Contribution

The results of this research can guide the design and implementation of biomedical question answering and other relationship extraction applications for precision medicine, precision oncology and other similar domains, where there is rapid emergence of novel knowledge. The methods developed and evaluated in this project can help NLP applications provide more accurate results by leveraging corpus based methods that are by design scalable and robust.

Precision oncology is rapidly evolving and scientists at cancer centers spend a significant amount of time and effort maintaining knowledge bases that directly affect clinical decision making processes.(Meric-Bernstam et al., 2013) As a preliminary step to this research, the AIMED project showed promising results in terms of helping expert curators find some of their desired answers in the literature. At the same time, AIMED also revealed some of the shortcomings of the Boolean retrieval system leveraging semantic constraints and co-occurrence frequency. The results of the current research are based on ranked retrieval by distributional techniques, and so, they are not directly comparable to the Boolean system of AIMED. Nonetheless, they elucidate the ways in which the applied models and configurations can be optimized to accommodate the unique characteristics of the problem domain of precision oncology.

In my final experiments in this project, I developed and evaluated a method in which cues were provided by NLP methods, without human intervention. This has important practical implications, as a data pipeline can be envisioned in which the initial selection and filtering

of relevant information from the literature is automated, which allows human experts to focus only on the information that has already been filtered, potentially saving time and effort.

It must be noted that the intended users for this system are annotators rather than clinicians. While the methods developed in the research have shown promising results, they are not yet at a level that can be used for direct clinical decision support without human supervision.

Future Steps

Future steps involve finding ways to improve the accuracy of my methods, test in other domains, and find ways to increase its practical usefulness.

There is great room for improvement in terms of increasing the accuracy of the results by developing methods that can incorporate more knowledge sources (like clinical trials, commercial drug company web sites, drug pipelines, etc.) to increase the breadth of available information. Both my preliminary work (AIMED) and the main research relied on ontology based named entity recognition (NER), using MetaMap (Aronson, n.d.). This approach posed limitations in terms of the breadth of the supported vocabulary, and as such, application of more accurate NER technology that is already available (Leaman & Gonzalez, 2008; Leser & Hakenberg, 2005) is a priority. Exploring other informatics approaches to build the search space, such as using dependency paths (as explained in the work by Percha (Percha & Altman, 2015)) to define relationships is another area of future research. A more comprehensive experiment is needed to determine the optimal set of

search space parameters, to accommodate the unique characteristics of rapidly evolving domains. The methods discussed in this dissertation have only been applied to the domain of precision oncology. Future work involves testing the techniques in similar domain where knowledge is rapidly evolving. To increase the practical usefulness of the system, the development of an interface to permit users to adjust query constraints in accordance with their preferences concerning workload and completeness, is an important step toward improving the system usability.

Conclusion

In this research, I compared relational to attributional measures of similarity across a range of representational approaches, for their ability to recover therapeutically important drug-gene relationships. Relational similarity performed better than attributional similarity for this task, demonstrating its utility as a means to identify clinically important biomedical relationships. These results have implications for the application domain of precision oncology, as they provide validation for methods that identify clinically-relevant drug/gene relationships. Furthermore, these methods should be applicable to the identification of biomedical relationships of any type where exemplar cues are available to seed the analogical retrieval process.

References

Ahlers, C. B., Fiszman, M., Demner-Fushman, D., Lang, F.-M., & Rindflesch, T. C. (2007). Extracting semantic predications from Medline citations for pharmacogenomics. In *Pacific Symposium on Biocomputing* (Vol. 12, pp. 209–220). Retrieved from https://books.google.com/books?hl=en&lr=&id=ytAaZYaW-k8C&oi=fnd&pg=PA209&dq=extracting+semantic+predications+from+medline+citation+for+pharmacogenomics&ots=waLQYfzM-_&sig=q1TqF_HzmOkDUz2-kVp820NtlXA

Apache Lucene. (n.d.). Retrieved April 12, 2018, from http://lucene.apache.org/

Aronson, A. R. (n.d.). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Retrieved from http://skr.nlm.nih.gov/papers/references/metamap_01AMIA.pdf

At, M. (1989). The UMLS Semantic Network. *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 503–507.

Athenikos, S. J., & Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, *99*, 1–24.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 238–247).

Basili, R., Hansen, D. H., Paggio, P., Pazienza, M. T., & Zanzotto, F. M. (n.d.). Ontological Resources and Question Answering. In *Workshop on Pragmatics of Question Answering, held in conjunction with NAACL 2004* (pp. 1–8). Retrieved from http://forskningsbasen.deff.dk/Share.external?sp=S3f703020-0197-11de-b05e-000ea68e967b&sp=Sku

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl 1), D267–D270.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., & Savova, G. K. (2011). The MiPACQ Clinical Question Answering System. *AMIA Annual Symposium Proceedings*, *2011*, 171–180.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., … Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, *44*(2), 277–288. https://doi.org/10.1016/j.jbi.2011.01.004

Chambliss, M. L., & Conley, J. (1996). Answering clinical questions. *The Journal of Family Practice*, *43*(2), 140–144.

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (pp. 166–174).

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, *6*(1), 57–71. https://doi.org/10.1093/bib/6.1.57

Cohen, T., Widdows, D., Stephan, C., Zinner, R., Kim, J., Rindflesch, T., & Davies, P. (2014). Predicting High-Throughput Screening Results With Scalable Literature-Based Discovery Methods. *CPT: Pharmacometrics & Systems Pharmacology*, *3*(10), 1–9.

Cohen, Trevor, Schvaneveldt, R. W., & Rindflesch, T. C. (2009). Predication-based semantic indexing: Permutations as a means to encode predications in semantic space. In *AMIA Annual Symposium Proceedings* (Vol. 2009, p. 114). American Medical Informatics Association.

Cohen, Trevor, Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, *43*(2), 240–256. https://doi.org/10.1016/j.jbi.2009.09.003

Cohen, Trevor, & Widdows, D. (2009). Empirical Distributional Semantics: Methods and Biomedical Applications. *Journal of Biomedical Informatics*, *42*(2), 390–405. https://doi.org/10.1016/j.jbi.2009.02.002

Cohen, Trevor, & Widdows, D. (2017). Embedding of semantic predications. *Journal of Biomedical Informatics*, *68*, 150–166.

Cohen, Trevor, Widdows, D., Schvaneveldt, R., & Rindflesch, T. C. (2011). Finding Schizophrenia's prozac emergent relational similarity in predication space. In *Quantum Interaction* (pp. 48–59). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-24971-6_6

Cohen, Trevor, Widdows, D., Schvaneveldt, R. W., Davies, P., & Rindflesch, T. C. (2012). Discovering discovery patterns with predication-based Semantic Indexing. *Journal of Biomedical Informatics*, *45*, 1049–1065.

Currie, L. M., Graham, M., Allen, M., Bakken, S., Patel, V., & Cimino, J. J. (2003). Clinical information needs in context: an observational study of clinicians while using a clinical information system. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 190). American Medical Informatics Association.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (n.d.). Indexing by Latent Semantic Analysis.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772. https://doi.org/10.1016/j.jbi.2009.08.007

Demner-Fushman, D., & Lin, J. (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, *33*(1), 63–103.

Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3519–3530).

Ely, J. W., Osheroff, J. A., Chambliss, M. L., Ebell, M. H., & Rosenbaum, M. E. (2005). Answering Physicians' Clinical Questions: Obstacles and Potential Solutions. *Journal of the American Medical Informatics Association : JAMIA*, *12*(2), 217–224. https://doi.org/10.1197/jamia.M1608

Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., & Evans, E. R. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ : British Medical Journal*, *319*(7206), 358–361.

Ely, J. W., Osheroff, J. A., Ebell, M. H., Chambliss, M. L., Vinson, D. C., Stevermer, J. J., & Pifer, E. A. (2002). Obstacles to answering doctors' questions about patient care with evidence: qualitative study. *Bmj*, *324*(7339), 710.

Fathiamini, S., Johnson, A. M., Zeng, J., Araya, A., Holla, V., Bailey, A. M., … Cohen, T. (2016). Automated identification of molecular effects of drugs (AIMED). *Journal of the American Medical Informatics Association*, *23*(4), 758–765. https://doi.org/10.1093/jamia/ocw030

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *BIOINFORMATICS*, *17*(1), S74–S82.

Fundel, K., Küffner, R., & Zimmer, R. (2006). RelEx-Relation extraction using dependency parse trees. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.2426&rep=rep1&type=pdf

Garraway, L. A., Verweij, J., Ballman, K. V., & others. (2013). Precision oncology: An overview. *Journal of Clinical Oncology*, *31*(15), 1803–1805.

GENTNER, D. (1988). Evidence for relational selectivity in the interpretation of analogy and metaphor. *The Psychology of Learning and Motivation*, *22*, 307–358.

Goodman, N. (1972). Seven strictures on similarity.

Gordon, M. D., & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, *49*(8), 674–685.

Gorman, P. N., & Helfand, M. (1995). Information Seeking in Primary Care How Physicians Choose Which Clinical Questions to Pursue and Which to Leave Unanswered. *Medical Decision Making*, *15*(2), 113–119. https://doi.org/10.1177/0272989X9501500203

Griffith, M., Griffith, O. L., Coffman, A. C., Weible, J. V., McMichael, J. F., Spies, N. C., … Wilson, R. K. (2013). DGIdb - Mining the druggable genome. *Nature Methods*, *10*(12). https://doi.org/10.1038/nmeth.2689

Hakenberg, J., Leaman, R., Vo, N. H., Jonnalagadda, S., Sullivan, R., Miller, C., … Gonzalez, G. (2010). Efficient Extraction of Protein-Protein Interactions from Full-Text Articles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, (3), 481–494.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Haynes, R. B., McKibbon, K. A., Walker, C. J., Ryan, N., Fitzgerald, D., & Ramsden, M. F. (1990). Online access to MEDLINE in clinical settings. A study of use and usefulness. *Annals of Internal Medicine*, *112*(1), 78–84.

Hersh, W., Cohen, A. M., Ruslen, L., & Roberts, P. (2007). TREC 2007 Genomics Track Overview. Retrieved from http://trec.nist.gov/pubs/trec16/t16_proceedings.html

Hersh, W. R., Crabtree, M. K., Hickam, D. H., Sacherek, L., Friedman, C. P., Tidmarsh, P., … Kraemer, D. (2002). Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *Journal of the American Medical Informatics Association*, *9*(3), 283–293.

Hersh, W. R., & SpringerLink (Online service). (2009). *Information retrieval: a health and biomedical perspective* (3rd ed.). New York, NY: Springer.

Herskovic, J. R., Tanaka, L. Y., Hersh, W., & Bernstam, E. V. (2007). A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association*, *14*(2), 212–220.

Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, *7*(04), 275–300.

HOLYOAK, K. J., & THAGARD, P. (1989). Analogical Mapping by Constraint Satisfaction. *COGNITIVE SCIENCE*, *13*, 29–5.

Hristovski, D., Dinevski, D., Kastrin, A., & Rindflesch, T. C. (2015). Biomedical question answering using semantic relations. *BMC Bioinformatics*, *16*(1). https://doi.org/10.1186/s12859-014-0365-3

Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of PICO as a Knowledge Representation for Clinical Questions. *AMIA Annual Symposium Proceedings, 2006*, 359–363.

Ide, N. C., Loane, R. F., & Demner-Fushman, D. (2007). Essie: A Concept-based Search Engine for Structured Biomedical Text. *Journal of the American Medical Informatics Association : JAMIA*, *14*(3), 253–263. https://doi.org/10.1197/jamia.M2233

Interactive MetaMap. (n.d.). Retrieved July 1, 2018, from https://ii.nlm.nih.gov/Interactive/UTS_Required/metamap.shtml

Interactive SemRep. (n.d.). Retrieved July 1, 2018, from https://ii.nlm.nih.gov/Interactive/UTS_Required/semrep.shtml

Johnson, A., Zeng, J., Bailey, A. M., Holla, V., Litzenburger, B., Lara-Guerra, H., … Meric-Bernstam, F. (2015). The right drugs at the right time for the right patient: the MD Anderson precision oncology decision support platform. *Drug Discovery Today*, *20*(12), 1433–1438.

Kanerva, P. (2010). What We Mean When We Say" What's the Dollar of Mexico?": Prototypes and Mapping in Concept Space. In *2010 AAAI Fall Symposium Series*. Retrieved from http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2243

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. In *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.6523

Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D., Ripple, A., & Rindflesch, T. C. (2008). Semantic MEDLINE: a web application for managing the results of PubMed Searches. In *Proceedings of the third international symposium for semantic mining in biomedicine* (Vol. 2008, pp. 69–76). Citeseer.

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, *28*(23), 3158–3160.

Kim, J.-D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *BIOINFORMATICS*, *19*(1), i180–i182.

Kotecki, M., & Cochran, B. (2002). Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relationsa J. Pustejovsky, J. Castaño, J. Zhang Department of Computer Science, Brandeis University 415 South St., Waltham, MA 02454, USA. In *Pacific Symposium on Biocomputing* (Vol. 7, pp. 362–373). Retrieved from http://psb.stanford.edu/psb-online/proceedings/psb2002/pustejovsk.pdf

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Biocomputing 2008* (pp. 652–663). World Scientific.

Lee, M., Cimino, J., Zhu, H. R., Sable, C., Shanker, V., Ely, J., & Yu, H. (2006). Beyond information retrieval—medical question answering. In *AMIA annual symposium proceedings* (Vol. 2006, p. 469). American Medical Informatics Association. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839371/

Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, *6*(4), 357–369.

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *ACL (2)* (pp. 302–308). Retrieved from http://www.aclweb.org/anthology/P/P14/P14-2050.pdf

Levy, O., Goldberg, Y., Dagan, I., & Ramat-Gan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Lin, D., & Pantel, P. (2001). DIRT@ SBT@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 323–328). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=502559

Lopez, V., Motta, E., Uren, V., & Sabou, M. (2007). *State of the art on Semantic Question Answering*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.107.1922&rep=rep1&type=pdf

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model of semantic representation. In *Brain and Cognition* (Vol. 30, pp. 5–5). ACADEMIC

PRESS INC JNL-COMP SUBSCRIPTIONS 525 B ST, STE 1900, SAN DIEGO, CA 92101-4495.

LUND, K., & BURGESS, C. (n.d.). Producing high-dimensional semantic spaces from lexical co-occurrence. Retrieved from http://csee.essex.ac.uk/staff/poesio/LAC/LAC03-04/lund_burgess_96brmic.pdf

McCray, A. T., Srinivasan, S., & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 235–239.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., & White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 491–498). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1219901

McInnes, B., & Pedersen, T. (2017). Improving Correlation with Human Judgments by Integrating Semantic Similarity with Second–Order Vectors. *BioNLP 2017*, 107–116.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for Similarity. *Psychological Review*, *100*(2), 254–278.

MEDLINE Citation Counts by Year of Publication. (n.d.). Retrieved June 25, 2018, from http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J., & Mills, G. B. (2013). Building a Personalized Medicine Infrastructure at a Major Cancer Center. *JOURNAL OF CLINICAL*, *31*, 1849–1857.

MetaMap Data File Builder. (n.d.). Retrieved July 1, 2018, from https://metamap.nlm.nih.gov/DataFileBuilder.shtml

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv Preprint ArXiv:1301.3781*. Retrieved from http://seed.ucsd.edu/mediawiki/images/e/e3/Wordembeddings.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL* (pp. 746–751). Retrieved from http://www.aclweb.org/anthology/N13-1#page=784

MOLDOVAN, D., CA, M. P., HARABAGIU, S., & SURDEANU, M. (2003). Performance Issues and Error Analysis in an Open-Domain Question Answering System. *ACM Transactions on Information Systems*, *21*(2), 133–154.

My Cancer Genome, Genetically Informed Cancer Medicine. (n.d.). Retrieved September 12, 2015, from http://www.mycancergenome.org/

Newman-Griffis, D., Lai, A. M., & Fosler-Lussier, E. (2017). Insights into Analogy Completion from the Biomedical Domain. *ArXiv:1706.02241 [Cs]*. Retrieved from http://arxiv.org/abs/1706.02241

Pakhomov, S. V. S., Finley, G., McEwan, R., Wang, Y., & Melton, G. B. (2016). Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, *32*(23), 3635–3644. https://doi.org/10.1093/bioinformatics/btw529

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, *40*(3), 288–299.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Percha, B., & Altman, R. B. (2015). Learning the Structure of Biomedical Relationships from Unstructured Text. *PLoS Computational Biology*, *11*(7). Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517797/

Percha, B., Altman, R. B., & Wren, J. (2018a). A global network of biomedical relationships derived from text. *Bioinformatics*, *1*, 11.

Percha, B., Altman, R. B., & Wren, J. (2018b). A global network of biomedical relationships derived from text. *Bioinformatics*.

Rekapalli, H. K., Cohen, A. M., & Hersh, W. R. (2006). A comparative analysis of retrieval features used in the TREC 2006 Genomics Track passage retrieval task. In *AMIA... Annual Symposium proceedings/AMIA Symposium. AMIA Symposium* (pp. 620–624). Retrieved from http://europepmc.org/abstract/med/18693910

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1994). The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, *123*(3), A12–3.

Riedel, S., Yao, L., Marlin, B. M., & McCallum, A. (2013). Relation Extraction with Matrix Factorization and Universal Schemas. The Association for Computational Linguistics. Retrieved from http://discovery.ucl.ac.uk/1410837/

Rinaldi, F., Dowdall, J., & Schneider, G. (2004). Answering questions in the genomics domain. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.498.3835&rep=rep1&type=pdf

Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, *36*(6), 462–477.

Rindflesch, T. C., Kilicoglu, H., Fiszman, M., Rosemblat, G., Shin, D., Kilicoglu, H., … Shin, D. (2011). Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use*, *31*(1–2), 15–21.

Sahlgren, M. (2005). An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. Retrieved from http://www.citeulike.org/group/3795/article/2227659

Sahlgren, Magnus. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*.

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

semanticvectors: SemanticVectors creates semantic WordSpace models from free natural language text. (n.d.). (Version Head revision). semanticvectors. Retrieved from https://github.com/semanticvectors/semanticvectors [Head Revision] (Original work published 2015)

SemMedDB Info. (n.d.). Retrieved from http://skr3.nlm.nih.gov/SemMedDB/dbinfo.html

SemMedDB_UTH Database Outline. (n.d.). Retrieved August 11, 2015, from http://skr3.nlm.nih.gov/SemMedDB/index_uth.html

Shalev-Shwartz, S. (2011). Online Learning and Online Convex Optimization. *Machine Learning*, *4*(2), 107–194.

Shang, N., Xu, H., Rindflesch, T. C., & Cohen, T. (2014). Identifying plausible adverse drug reactions using knowledge extracted from the literature. *Journal of Biomedical Informatics*, *52*, 293–310.

Shekelle, P. G., Ortiz, E., Rhodes, S., Morton, S. C., Eccles, M. P., Grimshaw, J. M., & Woolf, S. H. (2001). Validity of the Agency for Healthcare Research and Quality clinical practice guidelines: how quickly do guidelines become outdated? *Jama*, *286*(12), 1461–1467.

SNEIDERMAN, C. A., DEMNER-FUSHMAN, D., FISZMAN, M., IDE, N. C., & RINDFLESCH, T. C. (2007). Knowledge-based Methods to Help Clinicians Find Answers in MEDLINE. *J Am Med Inform Assoc*, *14*, 772–780.

The FDA's Drug Review Process: Ensuring Drugs Are Safe and Effective. (n.d.). Retrieved from http://www.fda.gov/drugs/resourcesforyou/consumers/ucm143534.htm

TREC Genomics Track. (n.d.). Retrieved August 7, 2015, from http://skynet.ohsu.edu/trec-gen

Turney, P. D. (1997). Similarity of Semantic Relations. *Computational Linguistics*, *1*(1).

Turney, P. D. (2005). Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th international joint conference on Artificial intelligence* (pp. 1136–1141). Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1642475

Turney, P. D., & Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*, *60*(1–3), 251–278.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Voorhees, E. M. (2001). Question Answering in TREC. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (pp. 535–537). New York, NY, USA: ACM. https://doi.org/10.1145/502585.502679

Westbrook, J. I., Coiera, E. W., & Gosling, A. S. (2005). Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association*, *12*(3), 315–321.

WESTBROOK, J. I., GOSLING, A. S., & PSYCHD, E. C. (2004). Do Clinicians Use Online Evidence to Support Patient Care? A Study of 55,000 Clinicians. *J Am Med Inform Assoc*, *11*, 113–120.

Widdows, D., & Cohen, T. (2015). Reasoning with vectors: A continuous model for fast robust inference. *Logic Journal of the IGPL*, *23*(2). Retrieved from http://www.oxfordjournals.org/our_journals/igpl/content/current

Widdows, D., & Cohen, T. (n.d.). The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. *Language*, *1*, 43.

Widdows, D., & Ferraro, K. (2008). Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. *LREC 2008*.

Widdows, D., Peters, S., Cederberg, S., Chan, C.-K., Steffen, D., & Buitelaar, P. (2003). Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13* (pp. 9–16). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1118960

Yu, H., & Cao, Y. (2008). Automatically Extracting Information Needs from Ad Hoc Clinical Questions. *AMIA Annual Symposium Proceedings, 2008*, 96–100.

Yu, H., & Sable, C. (n.d.). Being Erlang Shen: identifying answerable questions. Retrieved from http://cluster.cis.drexel.edu:8080/sofia/resources/QA.Data/PDF/M_2005_IJCAI_Yu_and_Sable_Being_Erlang_Shen--Identifying_Answerable_Questions-0520589825/M_2005_IJCAI_Yu_and_Sable_Being_Erlang_Shen--Identifying_Answerable_Questions.pdf

Yu, H., Sable, C., & Zhu, H. R. (n.d.). Classifying Medical Questions based on an Evidence Taxonomy. Retrieved from http://www1.cs.columbia.edu/~sable/research/aaai2005.pdf

Zweigenbaum, P. (2003). Question answering in biomedicine. In *Proceedings Workshop on Natural Language Processing for Question Answering, EACL* (Vol. 2005, pp. 1–4). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.9942&rep=rep1&type=pdf#page=10

Zweigenbaum, P. (2009). Knowledge and reasoning for medical question-answering. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions* (pp. 1–2). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1697289